2019

Echoes of Vision: Mental Imagery in the Human Brain

Jesse L. Breedlove
Medical University of South Carolina

Echoes of Vision: Mental Imagery in the Human Brain

by

Jesse Breedlove

A dissertation submitted to the faculty of the Medical University of South Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Graduate Studies.

Department of Neurosciences

2019

Approved by:

Chairman, Advisory Committee _____

Thomas Naselaris

_____

Truman Brown

_____

Jane Joseph

_____

Tom Jhou

_____

Andy Shih

Acknowledgements

I express thanks to my mentor, Thomas Naselaris from granting me the privilege to work on such an amazing project and for having faith in me to learn the skills that I needed to tackle it. Thank you for the scientific intuition you have imparted in me, and for inspiring me to not only be a thoughtful scientist but also to approach life with curiosity and an open mind. I also thank our postdoc, Ghislain St-Yves, for his guidance in upholding a healthy level of skepticism and in maintaining perspective in all aspects of life.

I would also like to express thanks to my committee members, Truman Brown, Jane Joseph, Andy Shih, and Tom Jhou for the varied perspectives they have lent to my research. I am incredibly honored that such an amazing group of scientists has taken sincere interest in my work and has devoted time to facilitate my progress. I thank Truman in particular for meeting my attitude and stubbornness with an equal amount of humor and wisdom. Thank you for instilling confidence in me and for fighting to provide me with the recognition that you believe I deserve.

I am grateful for my parents, Connie and John Droney, for all of the support and encouragement they have provided me over the years, and I thank my brilliant sisters, Mandie Breedlove-Hight and Alex Droney, for inspiring me to be a better person. I give extra thanks to Mandie for her thoughtful comments on this dissertation, and also for making learning cool and exciting at an early age when the rest of the world was telling me otherwise.

I thank the countless friends that have held me up through this process. In particular I am indebted to my soulmate, Heather Reynolds-Burns, who has shown up for every crazy endeavor I have decided to pursue with confidence, love, and a sense of free-spiritedness in tow. I also give genuine thanks to my cat confidantes, Numpy Cat and Hushpup, for keeping me company on many long nights and for making me laugh every single day. Finally, I give special thanks to Logan Dowdle, my best friend and partner in graduate school and life, for lending me his imaging skills and tireless enthusiasm for science. But more importantly, I thank him for his support, strength, and never wavering belief that the way I think about the world is a useful one.

My success would not have been possible without all of you; it truly takes a village to raise a PhD.

Contents

List of Figures

List of Tables

Abstract

JESSE BREEDLOVE. Echoes of Vision: Mental Imagery in the Human Brain. (Under the direction of THOMAS NASELARIS).

When you picture the face of a friend or imagine your dream house, you are using the same parts of your brain that you use to see. How does the same system manage to both accurately analyze the world around it and synthesize visual experiences without any external input at all? We approach this question and others by extending the well-established theory that the human visual system embodies a probabilistic generative model of the visual world. That is, just as visual features co-occur with one another in the real world with a certain probability (the feature “tree” has a high probability of occurring with the feature “green”), so do the patterns of activity that encode those features in the brain. With such a joint probability distribution at its disposal, the brain can not only infer the cause of a given activity pattern on the retina (vision), but can also generate the probable visual consequence of an assumed or remembered cause (imagery).

The formulation of this model predicts that the encoding of imagined stimuli in low-level visual areas resemble the encoding of seen stimuli in higher areas. To test this prediction we developed imagery encoding models—a novel tool that reveals how the features of imagined stimuli are encoded in brain activity. We estimated imagery encoding models from brain activity measured while subjects imagined complex visual stimuli, and then compared these to visual encoding models estimated from a matched viewing experiment.

Consistent with our proposal, imagery encoding models revealed changes in spatial frequency tuning and receptive field properties that made early visual areas during imagery more functionally similar to higher visual areas during vision. Likewise, signal and noise properties of the voxel activation between vision and imagery favor the generative model interpretation.

Our results provide new evidence for an internal generative model of the visual world, while demonstrating that vision is just one of many possible forms of inference that this putative internal model may support.

Chapter 1 : Introduction

What better way to demonstrate the rich and complex internal environment of the brain than through the ability to experience a scene in its absence? While its exact nature and utility have been subjected to exhaustive debate (as will be discussed later), it is difficult to deny that the thing we call "mental imagery" is intimately intertwined with our thoughts and daily experiences. Imagery-like phenomena appear to be associated with a plethora of cognitive and perceptual processes including spatial navigation (Byrne et al. 2007), dreams (Horikawa et al. 2013), future planning (Szpunar et al. 2007), and language comprehension (Just et al. 2004). Imagery is also thought to be involved in rehearsal (Savaki & Raos 2019), such as in sports (Filgueiras et al. 2018), and is thought to be linked to creativity (Palmiero et al. 2015).

While mental imagery's connection to our normal, everyday experiences is in itself interesting, it is in the context of clinical pathologies that the importance of advancing our understanding of imagery's underlying mechanisms is most apparent. An abundance of research on this topic has demonstrated that the presence of invasive or otherwise dysfunctional mental images across neuropsychiatric disorders is staggering, showing up in almost every major recognized category including anxiety, mood, addiction, and psychotic disorders as well as eating disorders and degenerative diseases (Brewin et al. 2010; Hackmann & Holmes 2004; Holmes et al. 2019; Holmes & Mathews 2010).

The manner and extent in which mental images are involved varies across disorders. When they are excessively intrusive, unwanted, and pervasive they can be associated with a number of anxiety disorders (Hirsch & Holmes 2007). This is most in post-

traumatic stress disorder, or PTSD (Hackmann & Holmes 2004). A core symptom used to diagnose this disorder is the reliving of a traumatic event (American Psychiatric Association 2013), often through "flashbacks", which are invasive and upsetting mental images related to the original event. Not only are these perceptual experiences a sensitive and specific indicator of PTSD in trauma survivors (Duke et al. 2008), their features (such as a sense of happening "here and now") can be early predictors of the severity and continuation of the symptoms beyond the immediate aftermath of the event (Kleim et al. 2007; Michael et al. 2005). Intrusive mental imagery is more than just a hallmark symptom of PTSD; the re-experiencing of traumatic memories through flashbacks may actually serve to reinforce and maintain the chronic disorder (Brewin 2011).

While the presence of intrusive images is a feature most often associated with PTSD, they are also prevalent in obsessive compulsive disorder (OCD), occurring in up to 75-90% of patients (Moritz et al. 2018; Speckens et al. 2007; Lipton et al. 2010). In these cases, the characteristic obsessions of the disorder are accompanied by vivid and distressing images, such as seeing or feeling dirt on one's skin or disturbing images of harming a loved one (Moritz et al. 2018). The presence of these intrusive images is associated with heightened anxiety (Speckens et al. 2007), and the strength of the imagery is associated with increased compulsive behavior in an apparent attempt to neutralize the images, subsequently leading to more impairment in daily function (Moritz et al. 2018). Interestingly, the most dominant sense for intrusive imagery in OCD is visual (Moritz et al. 2018; Speckens et al. 2007).

Recurrent and unwanted mental images have also turned up in a number of other anxiety disorders. Examples include third-person images of oneself looking anxious and sweating in social anxiety disorder (Hirsch & Holmes 2007), picturing being trapped in an inescapable situation or location in agoraphobia (Day et al. 2004), or imagining oneself as deceased in health anxiety disorder (Wells & Hackmann 1993).

Maladapted mental imagery is also found in mood disorders, often manifesting as an imbalance in the emotional affect produced by images. Individuals with depression have an impaired ability to imagine positive future events while their ability to vividly imagine negative events remains intact. Moreover, the positive events that they are able to voluntarily invoke are associated with less positive feelings (Holmes, Lang, et al. 2008). Much like the anxiety disorders, mood disorders (including major depression and bipolar disorder) are accompanied by intrusive *involuntary* negative imagery as well (Myers et al. 2007; Gregory et al. 2010). Outside of depressive states, patients with bipolar disorder additionally experience intense *positive* imagery (Close et al. 2014) which is thought to exacerbate the mania associated with the disorder (Holmes, Geddes, et al. 2008). Moreover, it appears that mental imagery has an even more sinister role in mood disorders as it may facilitate suicide in depressed populations via habituation, planning, and rehearsal (Braithwaite et al. 2010; O'Connor et al. 2018; Crane et al. 2012).

A number of studies have demonstrated a role of mental imagery in pathologies beyond mood and anxiety disorders. For example, intrusive mental images are also associated with eating disorders, such as body dysmorphic disorder, where the person has distorted and negative mental images of their bodily appearance (Osman et al. 2004).

Exceptionally vivid perception (MCGHIE & CHAPMAN 1961; Freedman 1974) and mental imagery accompany schizophrenia, the latter of which has been suggested as a trait marker of the disease (Oertel et al. 2009; Sack et al. 2005). A study investigating possible relationships between imagery and visual hallucinations in Parkinson's disease found that patients defined as "hallucinators" had stronger mental imagery than controls and the strength of imagery was correlated with degree of hallucinations (Muller et al. 2014).

The above mentioned research suggests that, in some cases, dysfunctional mental imagery is not only present but may also serve to maintain—or even play a causal role in— the disorder. While our appreciation of this is growing, an understanding of the underlying mechanisms of dysfunctional imagery is disproportionate to the clinical relevance. To capture how imagery gone awry in pathology we must first understand how it works in the healthy brain.

As is discussed in the next section, while we have made great strides in imagery research in the past few decades, we have, in some sense, uncovered more questions than we have answered. Consequently, our understanding of the generation, functional role, and even nature of normal imagery is still lacking.

Fortunately, recent advances in computational modeling offer renewed potential for imagery research, providing the framework and tools needed to build, test, and manipulate models of human systems (Kriegeskorte 2015; Kriegeskorte & Douglas 2018). Armed with these new tools, the current study seeks to understand how visual

mental imagery works in the healthy human brain; a task that, as it turn out, is in no sense a new endeavor.

History of Mental Imagery

Long-Standing Questions

Take a moment to imagine your favorite coffee mug. What color is it? Does it have a handle or any words written on the side? What are the relative proportions of its circumference to height? For a different example, consider the face of an analog clock at 6:50. Is the angle made by its hour and minute hands smaller or larger than 90 degrees?

In either case, was your experience at all *like seeing* your mug or a clock face? If so, did seeing the image in your "mind's eye" aid you in answering the questions? Incidentally, these are some of mental imagery's most ancient questions: 1) What is the *nature* of mental images? Are they, as the name implies, experienced as visual images? Or are they better characterized as language-like descriptions? And 2) Why do humans have mental images? Specifically, do they have any cognitive *utility*?

The various answers offered to these questions stretch back in time over two millennia and have substantially shaped the way we think about and research mental imagery today. Therefore, to understand the current landscape of imagery research and its gaps in knowledge, it is useful to consider the history of these questions and the philosophical and scientific research they sparked.

The first question concerning the nature of images in particular split all conversations about imagery into two views rather early on (MacKisack et al. 2016), a dichotomy that

has survived the intervening years. On one side, there are those who believe that mental images are experienced in the perceptual sense and contain *depictive* visual features. On the other are those who believe that they are better characterized as *descriptions*. The dichotomy eventually culminated in what was dubbed the "imagery debate" in the 20th century, and the defenders of the depictive and descriptive sides have been branded the "iconophiles" and "iconophobes", respectively. Divisions over the question of *utility* have been less clear-cut but are generally thought to follow the division over the nature of mental imagery: those who take mental images to be pictorial are naturally led to understand them as things that can be used to complete cognitive tasks, much like percepts can be examined and used, while most who believe that images do not contribute anything extra to cognition tend to align with the iconophobe stance (MacKisack et al. 2016; Thomas 2018).

In the following sections, I briefly discuss a few key actors in the history of philosophical research on imagery, and consider the answers they offered for one or both of these two main questions: the question of the *nature* of mental images and the question of their *utility*.

I then transition into the scientific investigation of mental images, in particular within cognitive psychology and neuroscience, and consider what insight they have brought to these longstanding questions as well as new questions they have unearthed. Finally, I discuss what kind of model of imagery would be necessary to address these questions.

In the meantime, we start in ancient Greece, where the stage for the imagery debate was first set.

Mental images as depictions – the Iconophiles

The oldest surviving philosophical discussions on the *nature* of mental images comes from the classical Greek thinkers in the 4th century BCE. One of the most extensive writings on this comes down from Aristotle (Thomas 2018) who believed that sensation was the process through which forms, actively emanating from their objects, collide with and impress themselves upon the sensory organ of the observer, much like a stamp into wax. These impressions start internal movements which can later be re-instantiated in the absence of the object as mental images, or *phantasmata*. Aristotle therefore saw phantasmata as being necessarily like the sensory experiences from which they came: visual mental images were echoes of visual sensation. It was particularly on the point of "usefulness" that Aristotle deviated from his mentor, Plato. While Plato treated mental images as incidental and misleading counterfeits twice removed from the eternal forms (MacKisack et al. 2016), the Aristotelian perspective granted mental imagery a central role in human cognition, necessary for motivation, decision-making, communication, dreams, and even thought itself: "the soul never thinks without a phantasma" (Aristotle 1984).

Even as human understanding of anatomy and physiology evolved substantially (such as recognizing the brain as the center of mental processing rather than the heart), many upheld a similar depictive treatment of mental imagery. In the 18th century, Hume described mental images, or *ideas*, as faint versions of visual percepts, virtually identical

in content and character and differing only in their degree of intensity (Hume 2003). Nearly a century later, Descartes, following a similar vein and foreshadowing neuroscientific research to come, proposed a common brain substrate (albeit erroneously the pineal gland) for vision and imagery, whereby stored information about the percept could recreate patterns on this shared structure, bringing the experience back into consciousness (Thomas 2018).

The depictive stance was almost entirely silenced during a decades-long bout of behaviorism that denied imagery research scientific integrity altogether, but was resurrected in the late 1900s, this time with the support of psychophysical and, soon after, neuroscientific methods (MacKisack et al. 2016). The most prominent present day champion of the iconophile stance is Dr. Stephen Kosslyn, who developed a thorough theoretical model of mental imagery. According to this theory, feedback connections match stored memories with visual representations in lower-level areas so that, during imagery, the details of some recalled object can be realized in the "visual buffer" (i.e., the primary visual cortex). These visual details (which allow you to answer questions such as "are a German Shepherd's ears pointy?") are only implicit in the higher-level representation from which they were constructed: "because visual memories are stored in [an] abstract format during perception, in order to recall the local geometry of shape it is necessary to generate mental images in topographically organized areas" (Kosslyn et al. 2006, p.142). This generation of images in the retinotopic visual cortex affords them the same explicit spatial and otherwise depictive properties that percepts have. Consequently, Kosslyn's theory gives imagery special cognitive utility: images allow you to become

conscious of, interpret, and reason about things that are not currently available to your eyes (Kosslyn et al. 2006).

Mental images as descriptions – the iconophobes

Similarly distributed throughout imagery's history are arguments in favor of the descriptive, or "propositional", account of mental imagery. One such argument, made by Dennett (Dennett 1969) and Shorter (Shorter 2007), involves the issue of *indeterminacy*, a perceived contradiction in which one can imagine an object (such as a coffee mug) with certainty while simultaneously not being able to, or even needing to, imagine a given detail of that object (such as its color). Supporters of this argument posited that this paradox, along with others innate to the pictorial image, could be resolved by treating mental imagery as a "language-like" code that represents images much like words represent objects and ideas rather than depicting them (MacKisack et al. 2016). Dr. Zenon Pylyshyn, a major advocate of the propositional stance (and usually positioned in direct opposition to Kosslyn), further developed this idea by drawing from ideas emerging in computer science research, arguing that mental images are "more accurately referred to as symbolic descriptions than as images in the usual sense" (Pylyshyn 1973). Concerning the utility of images, Pylyshyn has also proposed a "null hypothesis", stating that there is nothing special in particular about mental imagery: "reasoning with mental images involves the same form of representation and the same processes as that of reasoning in general, except that the content or subject matter of thoughts experienced as images includes information about how things would look." (Pylyshyn 2002) –a hypothesis that he believes we have yet to reject.

Contemporary Scientific Findings

Although the history of philosophical research on imagery may have spanned at least two millennia (MacKisack et al. 2016; Thomas 2018), a scientific understanding of imagery has emerged only in the last several decades (Kosslyn & Thompson 2003; MacKisack et al. 2016; Crawford I.P. Winlove et al. 2018; Albright 2012). A major driver of this understanding has been the development of rigorous psychophysical methods for probing mental imagery (Podgorny & Shepard 1978; Kosslyn et al. 1978; Ishai & Sagi 1995; Pearson et al. 2008).

Many of these methods provided evidence for spatially structured representations of imagined objects and consequently have been used as arsenal by the iconophiles. A fundamental example of these cognitive studies involves having subjects judge whether two drawings of 3D geometric objects rotated relative to one another are the same 3D object or different. These studies found that the time it takes to make a decision corresponds linearly to the angle in which the object has to be rotated in order to be directly compared to the other (Shepard & Metzler 1971), suggesting that subjects were mentally rotating visual representations of the objects in their mind's eye. Others have found that concrete nouns that can have physical, visual representations such as "chair", are easier to remember than more abstract formless nouns such as "truth" (Paivio 1963). Taken together, these results suggest that mental imagery somehow serves a facilitative role in tasks such as learning and decision making (MacKisack et al. 2016).

Equally important has been the development of neuroimaging tools, particularly functional magnetic resonance imaging (fMRI), for noninvasively measuring brain

activity of humans engaged in mental imagery. A central finding of several decades of fMRI studies of imagery has been the confirmation of an extensive overlap between imagery and vision in the brain. At the level of brain activity, these studies have revealed that imagery engages the same brain areas that we use to see (Wheeler et al. 2000; Crawford I P Winlove et al. 2018), including even the primary visual cortex (Kosslyn & Thompson 2003). At the level of representation, studies have demonstrated the presence of correlated multivoxel activity patterns between vision and imagery (Stokes et al. 2009; Reddy et al. 2010; Cichy et al. 2012; Lee et al. 2012; Albers et al. 2013; Bosch et al. 2014). Importantly, the activity patterns generated during imagery encode the same kinds of low-level visual features encoded during vision, such as spatial frequency and retinotopic location (Slotnick & Thompson 2018), which can be used to decode imagined content (Naselaris et al. 2014; Thirion et al. 2006; Horikawa & Kamitani 2017).

New Questions and the Echoes of Old

Considering the psychophysical and neuroimaging results discussed above, can we say that we have finally laid the age-old questions of nature and utility to rest? Many take the activation of early visual cortex and decoding of low-level features of the imagined images as solid evidence that the images themselves are depictive in nature and consider the debate a closed book (Kosslyn et al. 2006). However others remain unconvinced, pointing out that activity in lower-level visual areas, even if structured, does not in itself necessitate that mental imagery relies on actual visual images. It could instead be the case that their activation represents mere epiphenomenal side effects which are not used in the essential processing that gives rise to our imagery experiences and observed behaviour

(Pylyshyn 2003; Thomas 2018). Indeed, our ability to decode specific information from an area of the brain does not in itself demonstrate that the rest of the brain uses that information for the process in question.

On the question of utility, the opposing side also argues that we have yet to reject the null hypothesis, that there is no evidence that imagery is a special form of reasoning and therefore has not been proven to add anything special to cognition (Pylyshyn 2002). On the other hand, the cognitive and neuroimaging findings discussed above seem to provide an intuitive argument for how mental imagery could in fact be special: lower areas could be used to "fill in" details associated with and left unspecified by a recalled memory or generated scenario. However, this intuition has yet to be formalized. Our research efforts have therefore reframed and provided some insight into the questions of nature and utility but these questions remain open nonetheless.

The neuroscientific findings have also unearthed new, equally exciting questions that the field has only just begun to grapple with.

The first, which I will call the *synthesis* question, arises from the fact that the visual cortex is involved (whether you believe its involvement to be consequential or not) in both vision and imagery. *How does a single apparatus become involved in both analysis and synthesis?* Many agree that image generation must entail some sort of reversal of the visual cortex. Indeed, imaging studies indicate that imagery is *associated* with top-down activation from higher visual areas such as parietal and frontal cortices (Mechelli 2004; Stokes et al. 2009). However, an explicit model about how mental imagery arises in this

context or how the representations imposed on lower-level areas are computationally determined has yet to defined or tested.

The second question has been shaped by the historical dichotomy discussed above—most efforts to understand mental imagery have been aimed at testing the iconophile theory. The consequence has been extensive examination of how vision and imagery are *similar*. While there may be substantial overlap in the substrate and types of representations between vision and imagery, our subjective experiences at the very least lead us to understand that there must be *differences* in the specific states associated with the two. Therefore, fMRI studies have succeeded in putting an upper bound on how different seen and mental images can be, but have yet to actually reveal *how* and *why* seen and mental images differ.

The Model We Need

In the following section I consider what elements and assumptions a model of mental imagery would require to effectively address the outstanding questions detailed above. Specifically, I attempt to describe a model that can speak to why we have images and the divide over what they are like, as well as provide a story for how they are generated from the same substrate as vision while leading to two different phenomenological experiences. As I address these in turn, I add to the components that build up our theory.

Addressing the Nature Debate with Reinstatement in a Visual Hierarchy

Why has the imagery debate remained an open dispute for so long? Some have suggested that views were so strongly polarized because the opposing defenders

themselves experienced varying degrees of vividness in their own mental imagery (Reisberg et al. 2003).

Another (potentially related) possibility is that the debate was based on a false dichotomy whereby it was assumed that the two views were necessarily competing and mutually exclusive hypotheses on the nature of representations in the brain. This may stem from a separation in how we experience percepts and thoughts: while percepts feel concrete and external to oneself, thoughts are incorporeal and private. Perhaps when presented with something that is positioned in the undefined space between, we feel compelled to put it in either one bin or the other.

However, neurophysiological and imaging studies have now made it clear that the brain contains hierarchical structures (Markov et al. 2013; Piras et al. 2017). The visual system in particular contains multiple distinct areas that can represent the same external object at various, increasingly abstract levels. From a collection of edges all the way up to semantic categories and beyond, the visual system seems less of an isolated structure and more of a series of representations that runs seamlessly into the rest of the brain. Given this information, it becomes much less of a clear-cut task to draw a line in the brain where vision stops and non-visual thought begins. Therefore a mental image could occupy, just as percepts do, both a depictive representation and a higher-level "language-like" representation.

Here we present a model of mental imagery that emphasizes its representation across a hierarchy of visual structures. Therefore we start with multiple discrete levels arranged

in a chain (nodes in **Figure 1.1**), where each level represents the activity pattern in a given area of the brain encoding some set of features.

Interestingly, as Thomas (2018) points out, most on either side of the debate have assumed *intentionality*: "A mental image is always an image of something or other (whether real or unreal), in the same sense that perception (whether veridical or not) is always perception of something."

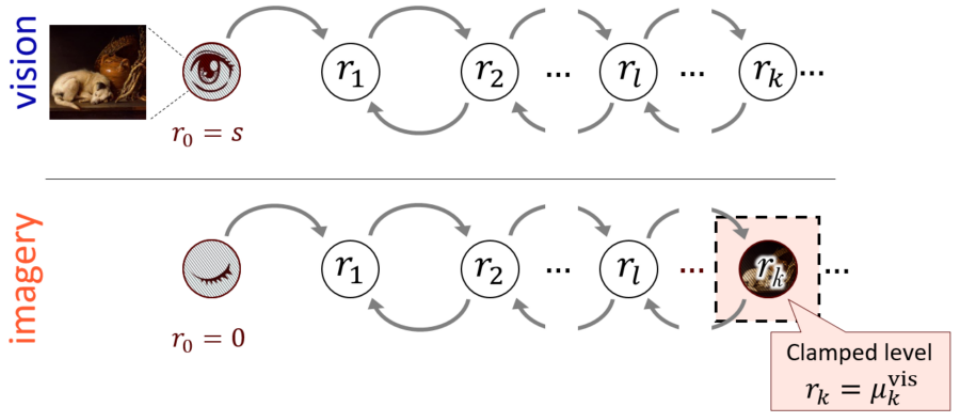

Figure 1.1 *Imagery as reinstatement in a hierarchical generative model*. Schematic showing vision and imagery in a recurrently connected (gray arrows) hierarchy of visual areas. Each node represents the activity pattern at a given level along the hierarchy while viewing (top) or imagining (bottom) some stimulus (*s*). When seeing *s*, the retina (r_0) is clamped to *s* while the remaining nodes converge to an activity pattern that encodes the causes of this pattern on the retina. When imagining *s*, the retina is clamped to an uninformative value (e.g. $r_0 = 0$) while at least one higher processing stage is clamped to the visual activity pattern that would be present when viewing *s* (i.e. reactivation; orange box). The remaining nodes below the clamped area converge to an activity that encodes the consequences of the cause specified by the reactivated level.

We take this shared intentionality to imply that there is shared representation at some level of the visual hierarchy. For example, whether one is seeing or imagining a zebra, one knows that the percept or image is *of* a zebra and not, say, a tiger. We might expect then that some area, perhaps at the level of processing responsible for encoding the abstract "zebra", converges during imagery to the activity it would be when seeing a zebra. Our model therefore assumes that during imagery there is a reinstatement of the visual activity pattern somewhere along the brain's hierarchy. Likely, this reinstatement occurs relatively high in the visual hierarchy given the high degree of similarity between vision and imagery in higher visual areas (Pearson et al. 2015; Dijkstra et al. 2019). We refer to the state of a processing level being held to a certain activity pattern as *clamping*

(denoted by the dashed orange box in **Figure 1.1**). Note that this just assumes that imagery involves reactivation at *some* point in the hierarchy. More notable is how the activations of the rest of the hierarchy are determined during imagery. Our proposed mechanism for this is discussed below.

Addressing the Question of Analysis/Synthesis with a Generative Model of Visual Perception

For our question of how the same system can both analyze retinal input (vision) and synthesize without input (imagery) we turn to a model of vision that is an alternative to the standard feed-forward discriminative model, one that posits that the brain must already contain the ability to produce images in order to *see*. Evidence for this has emerged primarily from studies designed to test, or derive the consequences of, the hypothesis that the visual system embodies a *generative model* of the visual world (Bar 2009; Friston 2005; Lee & Mumford 2003; Rao & Ballard 1999; Spratling 2016; Yuille & Kersten 2006). Generative models are systems of knowledge that support inference about what is uncertain given what is known (Christopher M. Bishop 2006). As the name suggests, the hallmark of a generative model of the visual world is the ability to *generate images*. In the context of vision, this means that the brain, with only input sensory nerve signals to go off of, infers what is out in the world by generating the causes of those signals. This is sometimes aptly referred to as "analysis by synthesis" (Yuille & Kersten 2006). A variety of phenomena in visual cortex such as spontaneous dynamics (Berkes et al. 2011), stimulus response non-linearities (Coen-Cagli et al. 2015; Karklin & Lewicki 2009; Rao & Ballard 1999), the encoding of prediction error (Murray et al. 2002; Alink et al. 2010), the structure of visual representations in low-level (Olshausen & Field 1996)

and high-level (Stansbury et al. 2013) visual areas, and the emergence of structured hallucinations as a consequence of damage to the visual system (Reichert et al. 2013) can be interpreted as evidence that the visual system embeds a generative model of the visual world and uses it to perform inference. While theories equating vision with inference in a generative model imply a compelling computational rationale for image generation in the visual system, an explicit hypothesis about how mental imagery arises in such a system has yet to be articulated or tested.

We adopt and adapt this theory to describe how vision and imagery might be derived from a single system. To expand briefly, the generative model (i.e. the visual system) contains a model of the outside visual world. That is, it replicates the statistical regularities found between different visual features in the sensory environment in terms of brain activity patterns that encode those features. Therefore, just as certain features are likely to co-occur in the external visual world (e.g., the feature "sky" and the feature "blue"), so too are the activity patterns that encode those features at different levels of the brain. In this model, *seeing* then entails using these relationships between features to infer the causes of the activity pattern on the retina.

What we propose is that imagery—like vision—is defined as inference in the same generative model. Note that the structure of the model, the hierarchy and the probabilistic relationships, remain constant. The only thing that changes is the inference that is made: the activity patterns of the multiple levels of the hierarchy during imagery (the unknown) is inferred from the activity of the clamped layer (the known) rather than the retina, which is uninformative. In other words, it can use the relationships specified by the

model to “unpack” the activations of lower areas that are likely given the activity at the clamped.

Within the framework of a visual hierarchy, bi-directional connections between levels (much like those found in the human brain) are needed in order to converge on mutually consistent representations across the network. Such connections are also necessary to infer activity in any lower level area given only the activity of a higher level (as is the case in imagery). Therefore we assume recurrent connections between each layer and the next that work to spread information started as activation in the clamped layer (whether at the retina or some level higher up) to the rest of the system (grey arrows in **Figure 1.1**).

Addressing the Question of Differences Through Independent Models of Imagined Features

The hierarchical generative model (HGM) theory of imagery that we propose makes specific predictions about how the activity patterns elicited by the two different types of inference will compare (see Chapter 2 for full treatment). Revealing any such differences would require directly comparing explicit, validated models of the features represented during imagery and vision of the same stimuli. Note that the model for imagery needs to be estimated directly from signals collected during imagery, independent of vision, in order to make a meaningful comparison. Previous studies investigating the kinds of features encoded during imagery have built models based on visual activity patterns and then tested these models on imagery activity patterns (Naselaris et al. 2014; Senden et al. 2019). While powerful, this tells you how well features encoded during vision can explain variance during imagery (i.e., how similar imagery is to vision) but falls short of

telling you to *what* imagery might otherwise be tuned. We therefore need to rely on a different experimental paradigm that can tell us what representations are activated by vision and imagery separately. As the next chapter will detail, the formalization of our theory predicts an explicit, testable relation between activity during imagery and vision and reveals a way in which we can directly model the types of features encoded during imagery.

Addressing the Question of Utility with Formal Inference

A computational model for mental imagery, like the one presented here, has the advantage over decoding alone in that it provides a testable story for why the brain would be encoding specific features in the areas that they are found. Therefore it has the potential to link findings that show visual features are encoded in primary visual areas during imagery to a theory of imagery utility. Our theory formalizes the intuitions of utility discussed earlier by equating the "reasoning" supported by mental imagery with inference in a generative model.

In summary, we propose that mental imagery is inference about the sensory consequences of predicted or remembered causes in an internal hierarchical generative model. The following chapters formalize this theory and describe how imagery imposes a different set of conditions than that of vision, leading to testable predictions about the types of features encoded during imagery.

Chapter 2 : A Formalized Theory of Mental Imagery: Inference in a Hierarchical Generative Model

In this chapter I further describe and formalize the unifying computational account of vision and imagery proposed above by expanding on an influential theory that describes vision as probabilistic inference in an internal hierarchical generative model (HGM). The structure of the HGM and the specifics of the particular formulation used here are detailed below.

Generative Model of Visual Cortex

As was introduced in Chapter 2, the "generative" portion of the HGM posits that the visual system performs a sort of "analysis by synthesis"; it makes sense of the stimulus-induced patterns on the retina by attempting to generate the *cause* of those patterns. Seeing therefore consists of activity patterns at discrete stages in the brain that encode different visual features of the proposed cause of the retinal activity. **Figure 2.1** illustrates this arrangement: $r_0, \ldots, r_L$ are the activity patterns of $L + 1$ stages in the visual system, where r_0 is the activity pattern of the retina and the remaining are the activity patterns in functionally distinct areas of the visual cortex (e.g. V1, V2, etc.). The probability of these activity patterns co-occurring in the network reflect the probability of the features they encode co-occurring in the visual environment. For example, the feature "sky" has a high probability of occurring with the feature "blue" (and less so with the feature "red"), so the activity pattern that encodes "sky" in one visual area has a high probability of occurring with the activity that encodes "blue" in another area. In this way

the brain contains a joint probability distribution for all the possible features at all stages, effectively amounting to an implicit model of the external sensory environment.

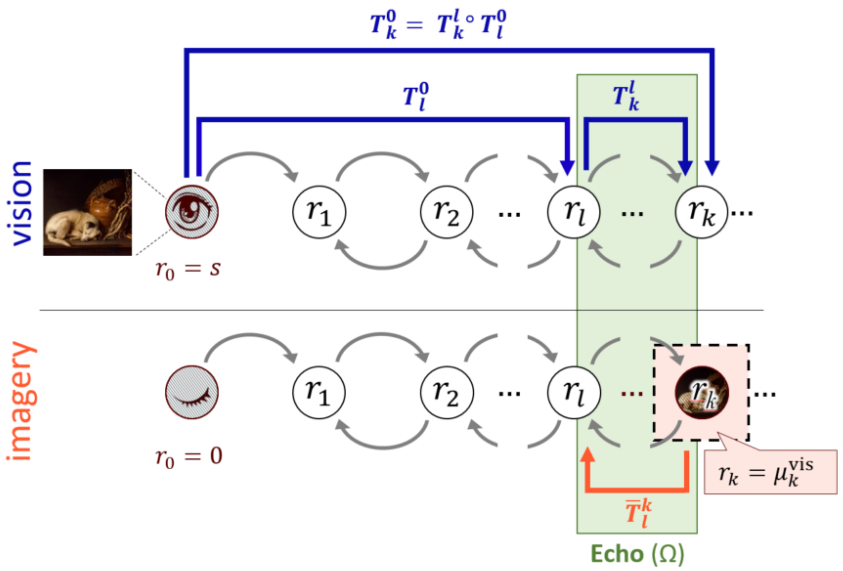

Figure 2.1 *Vision and imagery in a generative hierarchical network*. Schematic showing vision and imagery in a recurrently connected (gray arrows) hierarchical network. At equilibrium, the visual activity pattern at a processing stage, r_k, can be expressed as a transformation T_k^0(long blue arrow) of activity at the sensor stage r_0 or, equivalently, as a composition of transformations (shorter blue arrows) of activity patterns from lower stages. During imagery, $r_0 = 0$ and at least one higher processing stage is clamped to its visual activity pattern. Imagery activity patterns beneath the clamped stage (e.g., r_l) differ from their visual activity patterns by an echo **Ω**. The echo is a transformation from the current to the clamped layer (shortest blue arrow) and from the clamped layer back down (orange arrow).

Hierarchy of Causal Relationships

The model is "hierarchical" because the stages are arranged in a chain and the joint distribution described above is specified by a hierarchy of causal relationships. In other words, the activity pattern at the top of the hierarch r_L encodes features that, in the sensory environment, *cause* the features encoded at the stage below. For example, if r_L encodes object categories, then lower stages might encode visual features like color or texture that are caused by the presence of an encoded object category. If the object "zebra" causes the presence of the texture "stripes" in the sensory world, then the activity pattern in r_L that encodes "zebra" is likely to co-occur with the activity pattern in $r_{l<L}$ that encodes "stripes". Similarly, the specific activity pattern in r_L that encodes "building" would likely co-occur with an activity pattern in some layer $r_{l<L}$ that encodes vertical edges. Note that in the first example, it is "zebra" that is causally responsible for generating the feature "stripes" and not the other way around (it is the presence of a zebra that causes one to see stripes). By maintaining a representation of these causal relationships, the brain can "explain away" the lower level features by selecting the most probable cause of them. Also note that many different specific orientations and sizes of black and white stripes can be associated with a single object category "zebra", such that the specifics of the stripes can change over time (as they would if the zebra were to move about relative to the viewer) while the brain maintains that a zebra is the cause. This invariance of higher processing stages to changes in lower processing stages is critical to object identification. However, this also means there is a loss of resolution with ascension in the visual hierarchy. As will be discussed later, such an asymmetry in the structure of

the joint probability distribution leads to important differences in outcomes when the system performs inference under different conditions.

Formally, the hierarchical relationship between activity patterns is expressed by a set of conditional independence relationships between the various stages. Given an ordering $(0, \ldots, L)$ of activity patterns, the joint distribution can be expressed as a product of conditional distributions that specifies the interaction between one stage and the stage above it:

[1]

$$p(r_0, r_1, \ldots, r_L) = p(r_L) \prod_{l=0}^{L-1} p(r_l \,|r_{l+1})$$

In this hierarchical model, *seeing* a stimulus (s) means the network is conditioned on the activity in the bottom most layer (i.e. the retina, s_0) being set (or "clamped") to s (**Figure 2.1**, top). Vision is then the process of sampling activity patterns in all higher processing stages from the resulting posterior distribution $p(r_1, \ldots, r_L \,|r_0 = s)$, so that that the average activity state for a given visual area, l

[2]

$$\mu_l^{\text{vis}} = \mathbb{E}_{p(r_1, \ldots, r_L \,|r_0 = s)}[r_l]$$

encodes a feature that is the expected *cause* of the retinal stimulus s.

The HGM Expanded to Mental Imagery

We treat mental imagery as a subtly but importantly different conditional inference within the same HGM. Specifically, we propose that when a mental image of s occurs (**Figure 2.1**, bottom) the retina is clamped to some uninformative value (e.g., $s_0 = 0$) while the activity in some higher layer *k* is clamped to the expected activity pattern evoked by seeing s ($r_k = \mu_k^{\text{vis}}$; e.g. imagining a "house" might set some object recognition area within temporal lobe to the same activity that would be present there when seeing a "house"). Mental imagery is then the process by which the activity patterns in the remaining stages are sampled from the resulting conditional probability distribution $p(r_1, \ldots, r_{k-1},\ r_{k+1}, \ldots, r_L \big| r_0 = 0, r_k = \mu_k^{\text{vis}})$. The average activity pattern for the *l*th visual area during imagery

[3]

$$\mu_l^{\text{img}} = \mathbb{E}_{p(r_1,\ldots,r_{k-1},r_{k+1},\ldots,r_L \big| r_0=0, r_k=\mu_k^{\text{vis}})}[r_l]$$

therefore encodes an expected *consequence* of the cause specified by clamping the activity in a high-level visual area.

Vision and Imagery Encoding Models

For most HGMs it is not possible to explicitly write the conditional distribution without making assumptions about the distributions within the joint (e.g. Gaussianity), as the solution becomes computationally insurmountable when the variable being integrated

over has a large number of dimensions. This is especially so in our case where the posterior consists of the activity patterns of hundreds of thousands of voxels. However, note that during vision the only source of variance for each processing stage is the stimulus. Thus, the activity at each individual stage can be expressed as some function of *s* (i.e.f(s)).Therefore, without making any assumptions about the distributions, we can say that the expected activity pattern at l during vision, μ_l^{vis}, could be represented by some (possibly nonlinear) transformation, *T*, from the bottom stage 0 (the source of explainable variance), to the higher stage l. We write such a transformation of s as $T_l^0[s]$ where the superscript indicates which stage this function transforms *from* and the subscript indicates which stage it transforms *to*. We refer to any transformation of activity patterns from a lower to higher stage as a *forward transform*. Thus the expected activity pattern at l during *vision* can be expressed as

[4]

$$\mu_l^{\text{vis}} = T_l^0[s]$$

Incidentally, the expected activities during *imagery* can also be written as a transformation of activity patterns from the source of variance to the stage in question. Only now, as Equation 3 shows, the only source of variance in the hierarchy is the clamped stage which, when imagining s, equals the activity at that stage as it were when viewing s. Therefore the expected activity patterns during mental imagery can in general be expressed as

[5]

$$\mu_l^{\text{img}} = \overline{T}_l^k[\mu_k^{\text{vis}}]$$

for $l < k$, where $\overline{T}_l^k$ is a transformation of activity patterns from the clamped stage k down to stage l. We refer to any transformation of activity patterns from a higher to lower stage as a *backward transform* and denote it with a bar accent *($\overline{T}$)*.

Furthermore, in a strictly hierarchical architecture, the transformation from any one stage to another can be decomposed into transformations between intervening stages (**Figure 1.1**), e.g. $T_k^0 = T_k^l \circ T_l^0$ where $l < k$ is some intermediary stage between 0 and k. The imagery expected activities can therefore be rewritten as:

[6]

$$\mu_l^{\text{img}} = \overline{T}_l^k\left[T_k^0[s]\right] = \underbrace{\overline{T}_l^k \circ T_k^l}_{\Omega_{l,k}} \circ \underbrace{T_l^0[s]}_{E_l^{\text{vis}}[s]}$$

This equation and the underbrace notations highlight three major points that are useful for interpreting the relationship between vision and imagery as an experimentally testable prediction. First, we recognize the rightmost forward transform as a formal encoding model, denoted $\mathrm{E}_l^{\text{vis}}[s]$. In the context of visual neuroscience, an encoding model is a transformation of a visual stimulus into a prediction of evoked brain activity. A transformation that makes an adequate prediction of brain activity in a given portion of the brain (e.g. a voxel) serves as an indication of the kinds of information encoded in that portion. We refer to $\mathrm{E}_l^{\text{vis}}[s]$ as a *visual encoding model* as it transforms the stimulus s

into predicted activity patterns during vision. Thus we see that the expected activity pattern during mental imagery depends on the way the stimulus s is encoded *during vision*.

Secondly, the portion denoted by $\Omega_{l,k}$ indicates how the expected activity pattern during imagery will *differ* from the expected activity pattern during vision. Notice that $\Omega_{l,k}$ constitutes a forward transform from stage l to stage k, followed by a backward transform from stage k back to stage l (green box in **Figure 2.1**). It can thus be understood as an *echo* of the state of l during vision: imagery activity pattern at any one stage will resemble the visual activity pattern that has been fed forward from that stage to the clamped stage, then fed back to the original stage. The exact effect of the echo on the *types* of features encoded during imagery will be treated in detail in Chapter 5. Meanwhile, note that unless the feed-back transformation is an inverse of the feed-forward transformation the echo will induce a potentially measurable difference between the expected imagery activity pattern μ_l^{img} and the expected visual activity pattern μ_l^{vis}. In general, we would expect this difference to manifest as a loss of resolution in areas below the clamped area during imagery relative to the resolution normally seen in these areas during vision. This is due to the fact that lower areas during imagery are driven solely by the clamped area which, being higher in the visual hierarchy, has a lower inherent resolution compared to the retina. In other words, given the structure of the HGM proposed here, we expect lower areas during imagery to take on the lower-resolution properties seen in higher areas during vision. Note also that effect of the echo is compositional in nature, meaning that the distortion that each layer experiences relative

to its original representation during vision increases moving away from the clamped layer and toward the bottom layer.

This echo transformation therefore codifies the key signatures of inference we expect to observe during mental imagery: (1) an inheritance of functional properties from the reactivated brain area and (2) a gradient in this distortional effect that increases with hierarchical distance below the reactivated area.

Finally, the arguments above provide a guide for how to reveal the effects of the echo transformation on the features encoded in imagery activity patterns. Specifically, the composition of the two underbraced portions (the echo composed with the visual encoding model) defines something novel to the field: an *imagery encoding model*. Denoted as $\mathrm{E}_{l}^{\mathrm{img}}[s]$, the imagery encoding model is similar to its visual counterpart, only it predicts the activity pattern that will be evoked by *imagining* s. Since both the visual and imagery encoding models accept the same input s, it should be possible to explicitly characterize the differences between the encoding of imagined and seen stimuli by estimating imagery and visual encoding models.

In summary, we propose that imagery is the process of a particular form of conditional inference in the same HGM that allows us to see our external world. Formulation of this theory predicts measurable differences between the activity patterns encoding stimuli during vision and imagery. If the HGM is a good model for both forms of visual perception and such differences exist, we should be able to detect them by building

encoding models from brain activity measured during viewing and imagining the same stimuli.

Specific Aims

To this end, we have designed an experiment to capture brain activity from human subjects as they view and imagine matched visual stimuli. We have built and compared encoding models from the resulting activity patterns to test our hypothesized generative model of mental imagery. These objectives are summarized in the following aims.

Specific Aim 1: ***Build voxel-wise imagery encoding models.*** We hypothesized that imagery encoding models can be successfully measured from fMRI data, used to predict brain activity to new imagined stimuli, and used to decode the position and content of the imagined stimuli.

Specific Aim 2: ***Determine if signatures of inference in a generative model can be observed during mental imagery in the human brain.*** We hypothesized that (1) tuning to imagined features in lower visual areas will more closely resemble tuning to seen features in higher areas and (2) a gradient in this distortional effect that increases with hierarchical distance below the reactivated area, demonstrating key signatures of inference that follow from our formulation of mental imagery as inference in an internal generative model.

Chapter 3 : Experimental Methods

In this chapter, I describe in detail the experimental design, fMRI processing and display, training of the encoding models, as well as how specific measures of significance were determined. Some of the following methods will be reviewed again briefly in the following chapters as they become relevant.

Data Acquisition and Pre-processing

Subjects

Two healthy adult females and one healthy male participated in the main experiment, and one healthy male participated in the control experiment. All subjects had normal or corrected-to normal vision. All subjects gave written informed consent approved by the Institutional Review Boards at the University of Minnesota and/or the Medical University of South Carolina before participating in the study. Each subject completed both vision and imagery runs.

Experimental Design and Stimuli

The experimental scans were organized into separate 10-minute runs, each an uninterrupted succession of trials during which whole-brain blood-oxygen level dependent (BOLD) activity was measured. Runs were of two types: vision runs and imagery runs. During vision runs stimuli, including an object picture and a cue, were presented on a screen and viewed by the subjects. During imagery runs object pictures were not presented and only a cue was shown on the viewing screen. During these runs subjects instead imagined the cued object pictures (**Figure 3.1**). Subjects viewed the

stimulus on a 3M Vikuiti rigid rear projection screen projected on by a NEC NP4000 projector (1024 × 768 resolution and 60Hz). Data acquired during vision runs were used to estimate visual encoding models. Data acquired during imagery runs were used to independently estimate imagery encoding models. During all runs, subjects fixated on a 6-letter cue (filling a 1.5° × 0.4° rectangle) at the center of a grey stimulus field (16° × 16°). Eight brackets with 8 distinct colors framed the stimulus field throughout each run. Each bracket delineated a different but overlapping portion of the stimulus field (8° × 8°) within which an object picture might be seen (vision runs) or imagined (imagery runs). The same framing brackets were visible and unchanging at all times during all runs and conditions, and therefore contributed no variance in the stimulus. Cues were 6-letter descriptive abbreviations (e.g., “firtrk” cued a picture of a fire truck, “ababie” cued a picture of a baby) and always appeared at the same location and with the same dimensions throughout both run types (**Figure 3.2**).

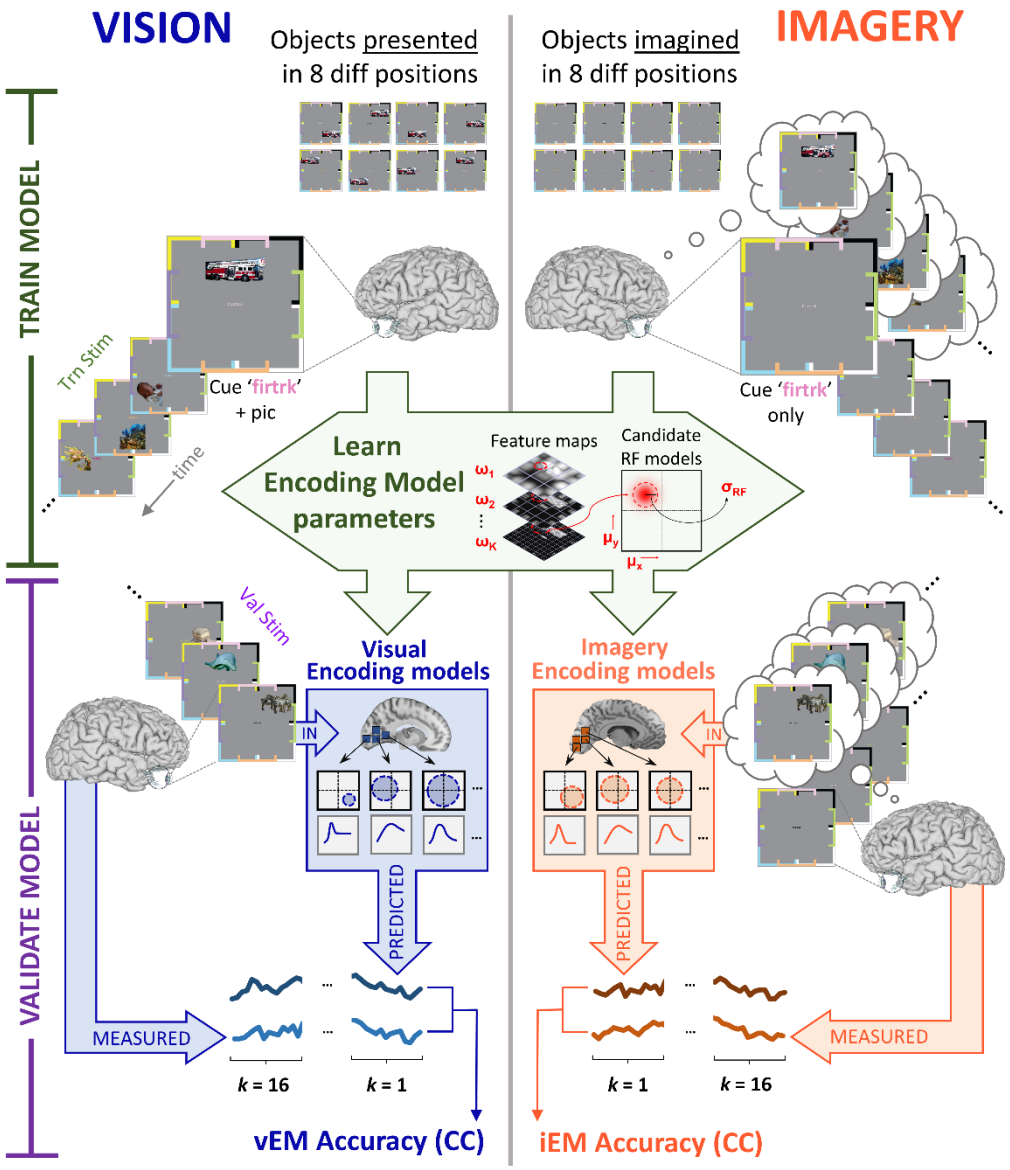

Figure 3.1 *Experimental Design.* Data and procedures for estimating visual encoding models (vEM, left) and imagery encoding models (iEM, right). Whole-brain fMRI (7T) measured BOLD activity as subjects viewed or imagined 64 unique object pictures at 8 distinct locations. The color of the six-letter cue for each stimulus coded a location bounded by a visible bracket. Model estimation (center) was applied separately to visual and imagery data, resulting in a distinct vEM and iEM for each voxel. Model prediction accuracy was k-fold cross-validated by computing Pearson correlation between predicted and measured activities on held-out data.

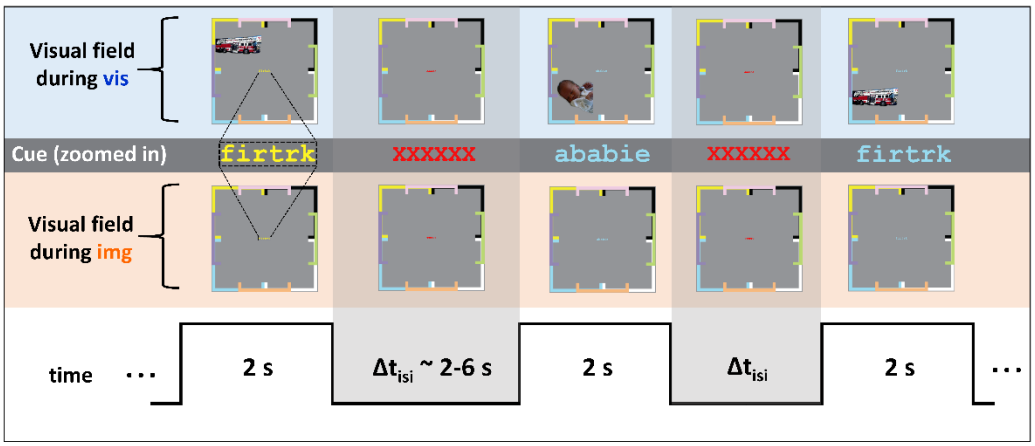

Figure 3.2 *Experimental Timing.* Top: The stimulus displayed on the viewing screen during vision runs. Second from top: Enlargements of the cues visible during both vision and imagery runs. Third from top: The display during imagery run. Bottom: Timing of stimulus on/off-set and inter-stimulus interval (Δt_{isi}).

During imagery runs (**Figure 3.1** right and **Figure 3.2** bottom), subjects were instructed to fixate on the cue and mentally project the cued object onto the portion of the visual field framed by the bracket whose color matched the color of the cue. For example, the cue "firtrk" written in yellow prompted the subject to imagine the firetruck picture in the upper left corner within the yellow framing brackets. Subjects were instructed to imagine the object in the correct position for the 2s duration in which the cue was present. Each imagined object was followed by an inter-stimulus interval (ISI) during which the 6-letter fixation cue at center was replaced by a dummy cue ("XXXXXX"). Subjects were instructed to stop imagining the object for the duration of the ISI which varied randomly from 1 TR (2 seconds) to $j \times TR$ where j was sampled from a Poisson distribution ($\lambda = 0.4$; $\approx$ 2 to 6s).

During the matched vision runs (**Figure 3.1** left and **Figure 3.2** top), subjects viewed that same display as during the imagery runs (i.e., the same cues, background, framing brackets, and ISI) except that the object picture was visually present. Subjects were instructed to fixate the cue and view the object pictures passively.

Object pictures (**Figure 3.3A**) were selected from the SUN labeled image collection (Xiao et al. 2010) and were selected to span the 19 object categories specified in (Naselaris et al. 2009). Each object was extracted from its background using the object mask provided by the SUN database. Masks were dilated by 10 pixels.

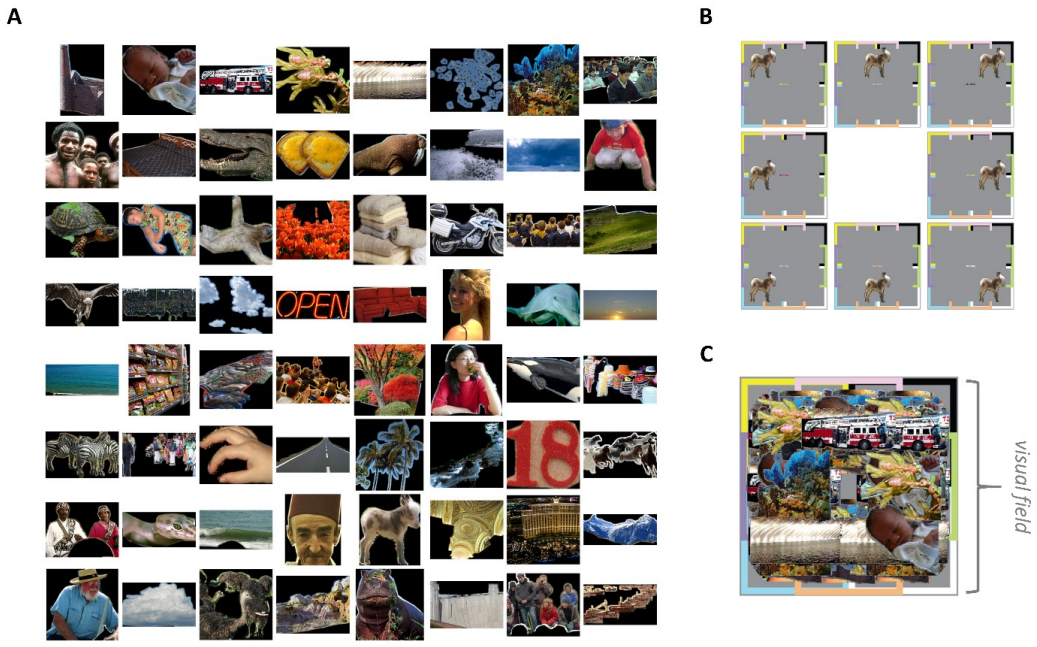

Figure 3.3 *Details of the stimuli.* **A)** All 64 individual object pictures viewed and imagined during the experiment. **B)** An object picture displayed in each of the 8 positions bounded by the framing brackets. **C)** A superposition of all 64 objects pictures showing the visual field coverage of the stimuli.

Eight unique object pictures (single row in **Figure 3.3A**) were displayed and imagined during each scanning session (set of runs). Note that subjects therefore had to remember only 8 object pictures at a time. Prior to each run subjects familiarized themselves with the experimental stimuli using a self-paced version of the imagery experiment. Familiarization sessions halted when subjects felt confident that the 8 object pictures and associated 6-letter cues were committed to memory. These sessions varied in duration from 20-30 minutes per scanning session.

Each object picture was displayed or imagined at each of the 8 framed locations (**Figure 3.3B**), for a total of 64 unique stimuli per run. Each unique stimuli was also viewed/imagined twice within a run, for a total 128 stimulus presentations/acts of imagery per run. Runs were repeated a minimum of two times for vision and two times for imagery. Repetition of runs allowed us to utilize a BOLD time-series denoising technique that uses the reoccurring conditions to cross validate parameter fits (Kay, Rokem, et al. 2013). Over all sessions, runs, object pictures, and positions there were a total of 512 unique stimuli (each viewed and imagined), at least 2048 stimulus presentations, and at least 2048 acts of mental imagery, per subject.

MR Acquisition Parameters

7T MRI data was acquired at the Center for Magnetic Resonance Research (CMRR) at the University of Minnesota. The experimental fMRI runs were collected using a 7T Siemens Magnetom scanner and a Nova Medical head coil (CP Transmit / 32 channel receive coil). Whole-brain functional data was acquired with a gradient-echo EPI sequence at a resolution of $1.6mm^3$: TR 2000 ms, TE 22.8 ms, FOV 130 × 130, Partial Fourier 7/8, 70 slices, GRAPPA R=2, multiband acceleration factor 2, anterior-posterior phase encode, transverse slice orientation.

3T MRI control data was acquired at the Center for Biomedical Imaging (CBI) at the Medical University of South Carolina. The experimental fMRI runs were collected using a 3T Siemens Trio TIM scanner and a 32-channel receiver coil array. Whole-brain functional data was acquired with a gradient-echo EPI sequence at a resolution of 2.5

mm^3: TR 2000 ms, TE 33 ms, FOV 100 x 100, 52 slices, multiband acceleration factor 2, anterior-posterior phase encode, transverse slice orientation.

Prior to experimental runs we collected a 1-mm T1-weighted whole-brain anatomical volume (at 7T for subjects in the main experiment and at 3T for the control subject). We also collected standard GRE fieldmaps at each scanning session for the correction of EPI spatial distortions (Jezzard & Balaban 1995).

Surface Reconstructions

Structural T1 volumes were skull-stripped and used to obtain surface reconstructions (Freesurfer). Flatmaps used for displaying results and drawing retinotopic ROIs were prepared with pycortex (Gao et al. 2015). Briefly, T1-weighted volumes were passed to Freesurfer's recon-all (version 6) for cortical reconstruction and segmentation, pial and white-matter surface rendering, and cortical inflation. We then made manual edits to the segmentations to ensure optimal surface quality. Digital cuts were made into the inflated surface using Blender (v2.78) and then processed by pycortex for flattening and rendering. Functional data to be displayed on surfaces were rigidly aligned to the above processed structural volumes using FSL FLIRT.

Functional Image Correction and Alignment

Functional scans were corrected and aligned within subject only. For each run, time series motion correction was performed through rigid alignment of all volumes to the middle volume (FSL MCFLIRT). Acquired fieldmaps were then used for spatial B0 distortion correction (FSL FUGUE). Functional volumes were temporally resampled to

correct for slice timing differences (FSL slicetimer). Spatial transformations up to this pre-processing stage were then concatenated and applied to un-corrected and un-registered volumes to minimize spatial resampling. An average of the time series from the run with the least amount of absolute movement was selected as the reference image for rigid alignment between runs (FSL FLIRT). Any residual misalignment was reduced via non-linear registration of all functional volumes to the same reference image (FSL FNIRT). Transforms for the last two registrations were concatenated and applied to the within-scan corrected images.

Time-Series Modeling, Denoising, and Activation Estimation

BOLD time-series modeling for each voxel in the corrected and registered functional volumes was performed using GLMdenoise (Kay, Rokem, et al. 2013) (Canonical HRF, visual cortex mask for PC-selecting voxels, noise-pool threshold defined as 99th percentile of R^2 values, minimum of 700 voxels with highest R^2 selected from visual cortex used to select number of principal components, 100 bootstrapping iterations). For each voxel this procedure output an estimate of activation amplitude per unique seen stimulus and an independent estimate of activation amplitude per unique imagined stimulus. Activation estimates were bootstrapped to obtain confidence intervals.

Region of Interest (ROI) Identification

We conducted independent retinotopic mapping experiments to identify visual areas V1, V2, V3, V3ab, V4, and LO. We utilized the mapping stimuli and population receptive field estimation (analyzePRF) technique from Kay et al. (Kay, Winawer, et al. 2013; Dumoulin & Wandell 2008) to construct angle and eccentricity maps for subjects

1, 3, and control subject 4. Similar retinotopic maps were constructed using a standard traveling wave approach for subject 2 (Engel et al. 1997). These maps were overlaid onto flattened cortical surfaces and imported into Inkscape where phase reversals and eccentricity patterns were used to hand-draw continuous ROI boundaries as described by Hansen et al. (Hansen et al. 2007). Ventral and dorsal regions were delineated for V1, V2, and V3. Surface-defined ROIs were then transformed back to functional 3D volumetric space using pycortex (get_roi_masks function with gm_sampler = "thick"). A cortical ribbon mask (adopted from Freesurfer's earlier segmentation) was prepared for all ROIs.

To identify cortex within the intraparietal sulcus (IPS) functional volumes were registered to Montreal Neurological Institute (MNI) $1mm^3$ standard space (FSL FNIRT, 3mm warp resolution). ROI's were then defined using published probabilistic maps of ROIs in volumetric standard space (Wang et al. 2015).These probabilistic maps were thresholded at 10%; any voxel belonging to multiple ROIs under this threshold was assigned to the ROI for which it had the highest probability of membership. The registration transform was then inverted to bring these ROIs from standard space back into individual subjects' native spaces.

Voxel-wise Encoding Model Design, Estimation, and Analyses of Parameters

The Feature-Weighted Receptive Field (fwRF) Encoding Model

Imagery encoding models and visual encoding models were estimated using the feature-weighted receptive field (fwRF) approach that was developed in-house (St-Yves

& Naselaris 2017). The fwRF is a voxel-wise encoding model meaning that it can be used to predict activation in response to arbitrary stimuli for each voxel in a functional volume (summarized in **Figure 3.4A**). Specifically, the fwRF utilizes both classic receptive fields and visual feature tuning by defining them as two separable sets of parameters of the same model. First, the model makes the assumption that a voxel's activity can be explained by multiple feature maps, each of which describes the degree to which a specific feature is present across the visual stimulus, preserving the topology of the visual field. Secondly, the model assumes that the region that a voxel responds to is relatively localized and fixed across all feature maps.

For this study feature maps were constructed by convolving Gabor wavelets of different sizes, orientations, and spatial frequencies with the visual stimuli (note that this preserves the topology but not necessarily the resolution of the stimulus in native visual space). Rather than pooling over pixels in a localized region of the stimulus (as the standard pRF model does) the fwRF model pools over the pixels in a localized region of feature maps (called the feature pooling field). Feature pooling in this study was adjusted such that the overall pooling was consistent across all feature maps (of different resolutions) with respect to a visual field region (the classical receptive field) (see **Figure 3.4B**). The fwRF objective is to minimize the squared error between the observed data and the prediction produced by a set of feature weights and receptive field. Feature weights were optimized for each voxel through stochastic gradient descent and RF size and locations were optimized through grid search.

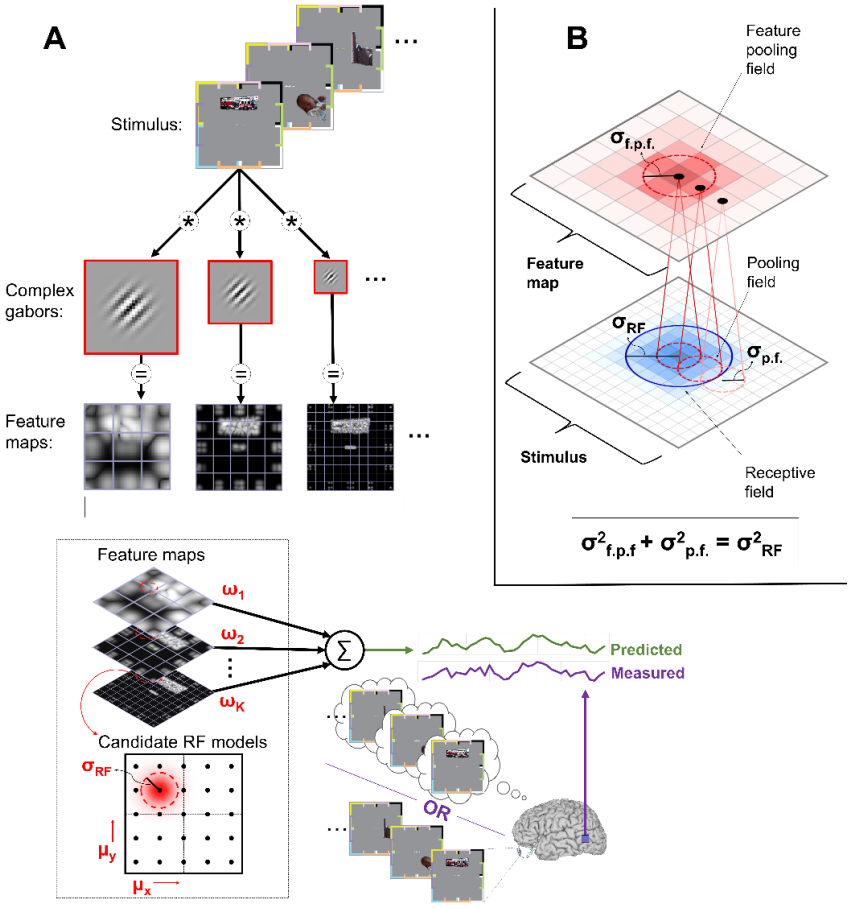

Figure 3.4 *The fwRF model.* **A)** Schematic illustration of fwRF model training. Training was performed independently for vision and imagery using completely independent datasets. First, the visual stimuli were convolved with complex Gabor wavelets of various spatial frequencies, orientations, and sizes. The visual stimuli used for both the imagery and visual encoding models was the same. Feature maps were constructed by taking the magnitude of each complex convolution. A feature pooling field (f.p.f) is then applied to each feature map. The f.p.f is a Gaussian function of space that is projected onto each feature map to obtain a scalar output per feature map. These outputs are then weighted by feature weights $(\omega_1, \ldots, \omega_n)$ that are, collectively, a visual feature tuning function. The sum of these feature-weighted outputs is the fwRF model's predicted activation in response to a stimulus. For each voxel, the sum of squared errors between the model's predictions and the measured activations in a training set was minimized using stochastic gradient descent over the feature map weights and a brute-force search over a grid of receptive field sizes (σ_{RF}) and locations (μ_x, μ_y). For the imagery runs, measured activations correspond to imagined stimuli; for the vision runs, measured activation correspond to seen stimuli. **B)** The feature pooling field size $\sigma_{f.p.f.}$ is constrained by the selected receptive field size σ_{RF} and the pooling field size $\sigma_{p.f.}$ of the feature being pooled over. As a consequence, feature maps with $\sigma_{p.f.} \geq \sigma_{RF}$ have zero weights.

For practical reasons, two versions of the fwRF were estimated for each subject: one "fine" model fit was performed on the voxels within labeled ROIs only (V ≈ 50K) while a "coarse" model was fit on whole-brain (all voxels available i.e. V ≈ 300K). The coarse fits were used to provide a brain-wide view of the model's prediction accuracy. The fine model fits offered more precision in model parameter estimates and slightly better prediction accuracy. Note that both models produced consistent results for the voxels to which they were both applied.

For the "fine" models the visual feature set consisted of Gabor wavelets at 4 uniformly distributed orientations times 12 spatial frequencies (log-spaced between ω = 0.35 and ω =11.0 deg^{-1}) for a total of 48 features. Gabor wavelets spanned 4 standard deviations of the Gaussian envelope and were designed to have one cycle per standard deviation. The fine model utilized a uniform grid of 21 × 21 receptive locations times 12 receptive field sizes (log-spaced between 0.22 and 8.75 degrees of visual angle).

For the "coarse" models, visual feature sets consisted of Gabor wavelets at 4 different uniformly distributed orientations times 8 spatial frequencies (log-spaced between ω = 0.96 and ω = 8.23 deg^{-1} for a total of 32 features. The coarse model utilized a uniform grid of 10 × 10 receptive locations times 6 receptive field sizes (log-spaced between 0.73 and 4.37 degrees of visual angle).

Training and Cross-Validation

The fwRF was applied to vision and imagery datasets independently, ultimately producing two encoding models for each voxel. Data from the visual runs were used to

estimate the visual encoding models, while data from the imagery runs were used to estimate the imagery encoding models. In both cases, the input of the encoding models s_t during training is the exact visual stimulus that was presented during the visual experiment (including the cue). All of the following details therefore apply to both instances of fwRF training.

Feature weights and receptive field parameters were estimated using a k-fold cross-validation procedure. To do this the $N_{\text{total}} = 512$ samples of activation per voxel (corresponding to the 512 unique object-position pairs) were split into k=16 randomly selected and nonoverlapping validation subsets ($N_{\text{val}} = 32$). For each fold of cross-validation, one of these subsets was set aside for validation while the rest (k-1) were used as model-training sets to estimate a fwRF model for each voxel (the latter group was further broken up into training and hold out sets to prevent overfitting). The resulting fwRF models were used to predict the activation of each voxel in response to the held-out validation subset. This process was repeated k times so that there was a prediction of activation for all voxels in response to each of the 512 unique object-position pairs. For the "fine" model the process was repeated $k = 16$ times, resulting in 16 distinct fwRF models for each voxel. For the "coarse" model the process was repeated $k = 8$ times. For each voxel, the concatenated 512 activation predictions were compared to the corresponding measured activation from the brain to obtain a Pearson correlation coefficient characterizing the overall model prediction accuracy for this voxel (see bottom of **Figure 3.1**). Error estimates on prediction accuracy values were obtained by

sampling 100 times with replacement the k models for each voxel and recalculating the correlation coefficient for each sample.

Estimation of the feature weights for each model was performed using stochastic gradient descent with a learning rate of 5×10^{-3} and with a batch size of 96 for a maximum of 100 epochs. For the "fine" model 40% of the training data was held-out as an early stopping set; for the "coarse" model the hold-out was 50%. Parameter updating halted early if the held-out loss began to increase. Estimation of the receptive field location and size was performed by brute-force search over the minimum hold-out loss reached by all possible candidates on a grid (see **Figure 3.4A**).

A significance threshold on prediction accuracy (dashed grey lines in **Figure 5.2**A and **Figure 5.3**) was defined as three standard deviations ($p < 0.01$) from the mean of a null distribution over prediction accuracy that assumed no relationship between the model predictions and measured activities. This null distribution was built through 500 iterations of shuffling the model's predicted activity over conditions for each voxel and then measuring the correlation coefficient between this shuffled predicted activity and the corresponding measured activity for that voxel. Unless otherwise specified, analyses of receptive field attributes and spatial frequency tuning were applied only to voxels with a visual or imagery encoding model above this accuracy threshold (Pearson correlation coefficient $\geq$ .16).

Voxel Selection

To ensure that our results did not reflect any response to the slight changes in hue and shape of the cue with condition, the following procedure was used to identify and discard any voxel that showed sensitivity to the 6-letter cues during either the imagery or the visual runs. The visual stimuli used to estimate each fwRF model included both the object picture and its associated cue. To test for sensitivity to the cue, we calculated cross-validated prediction accuracy using input stimuli that contained either the cues or the object pictures only. Voxels for which the cue-only stimuli resulted in above-threshold prediction accuracy for either the imagery or the vision encoding model were discarded from any analysis of receptive field attributes of feature tuning. Only voxels for which the picture-only stimuli resulted in above-threshold prediction accuracy were retained for receptive field and feature tuning analyses. **Error! Reference source not found.** enumerates the number of voxels per ROI and subject that satisfied these conditions. In **Figure 5.4**B, we retain voxels for which either the visual (top) or the imagery (bottom) encoding models had above-threshold predication accuracy. In **Figure 5.4**C-E we retain only voxels for which both the visual and imagery encoding models had above-threshold prediction accuracy.

Figure 3.5 shows receptive fields of the discarded cue-responsive voxels. As expected, receptive fields tend be small and are concentrated at the center of the visual field where the cue was displayed.

Table 3.1 *Number of selected voxels per ROI per subject*. Each cell contains three rows of numbers. The top row gives the total number of voxels in the specified ROI. The middle row gives the number of voxels with a visual encoding model that exceeds the prediction accuracy threshold (left, blue) and the number of voxels with an imagery encoding model that exceeds threshold (right, orange). The bottom row (dark green) gives the number of voxels with a visual and imagery encoding model that both exceed the prediction accuracy threshold.

	V1	V2	V3	V4	LO	V3a/b	IPS
S1	3567 1387 113 87	4371 1758 187 162	4527 1826 450 431	3305 1283 404 388	3610 1112 318 283	1970 891 565 507	15172 847 496 306
S2	3168 1136 214 110	3282 1060 358 216	2695 849 595 393	1768 358 274 154	1964 716 433 370	1459 616 432 366	15242 954 1030 536
S3	3249 959 106 97	3278 929 224 150	3414 972 250 184	2201 506 83 63	1439 277 123 94	1707 443 147 119	15090 219 190 40

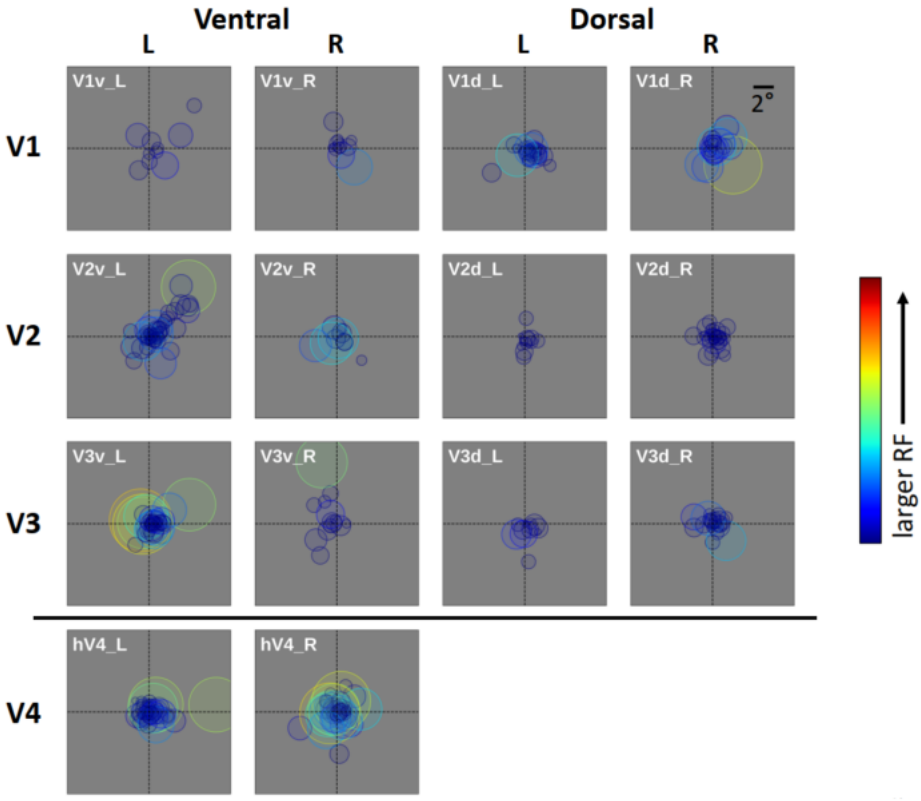

Figure 3.5 *Receptive fields of removed voxels with cue responsivity*. Circle plots showing the average raw receptive fields (locations and relative sizes) corresponding to voxels from each area that were well predicted by the cue and subsequently removed from further analyses. Each ROI (row) is partitioned by hemisphere ("L" = left and "R" = right), and/or dorsal ("d") and ventral ("v"). Circle radius is one standard deviation of the corresponding Gaussian envelopes. As expected, the discarded voxels had small receptive fields centered on the location of the cue.

Receptive Field Size and Location

As described above, for each voxel we fit k = 16 independent visual encoding models, and k = 16 independent imagery encoding models, each corresponding to a different training/validation split of the data. Thus, for each voxel we obtain 16 different estimates of receptive field size and location. Results on differences in the location **Figure 5.6C-D** and size **Figure 5.6A-B** between imagery and visual receptive were obtained by repeatedly sampling these estimates.

To construct the plots in **Figure 5.6C-D** we sampled 1000 pairs of imagery and visual receptive field parameters at random from the k = 16 "fine" encoding models available for each voxel. Lines in **Figure 5.6D** show the average shift (over all samples) in receptive field location of individual voxels from vision to imagery. Values in **Figure 5.6**C show the average over all sampled pairs and voxels in each ROI, and the error bars for each subject indicate one standard deviation of the sampling distribution and the yellow shading shows the same for combined subject data. Similarly **Figure 5.6B** shows the mean and one standard deviation of the differences between imagery and visual receptive field sizes. Illustrations of receptive fields in **Figure 5.6A** and **Figure 3.5** show average receptive field locations and sizes over samples for individual voxels.

To obtain the significance estimates displayed for all receptive field size and location results in **Figure 5.6** we tested the hypothesis of a non-zero mean difference between imagery and visual receptive field parameters against the null hypothesis of no mean difference. To construct the null distribution we performed the same sampling process as above with the addition of randomly assigning the "imagery" or "vision" designation to

each sampled value for each voxel in the indicated ROI. We then calculated the mean difference between the receptive field parameters between each group. This process was repeated 1000 times, resulting in a histogram of differences in mean receptive field parameters for each ROI. The region outside the red shaded area in **Figure 5.6B-C** indicates values at significance level $p < 0.01$ for combined subject data. For each subject and point individually, the observed values with $p < 0.01$ are indicated by a black asterisk on the mean observed value.

Spatial Frequency Tuning

Tuning is a reflection of the relative preference of a voxel for a certain type of explanation (i.e. certain parts of the model). For example, if a voxel's model was to crucially dependent on a certain set of features such that its prediction dropped significantly whenever the weights associated with those features were set to zero, then we would say that the voxel is tuned to these features. Such a dropout procedure (St-Yves & Naselaris 2017) was used to determine all spatial frequency tuning shown and analyzed in this study. To determine tuning for a given voxel to a specific frequency, we set all weights to 0 except those belonging to feature maps generated using Gabor wavelets of that frequency. We then calculated the Pearson correlation ρ between the activation predicted by the model with only those weights and the measured activation. The value of ρ was calculated for each spatial frequency and can be interpreted as a percentage of variance explained (David & Gallant 2005) by that frequency. In order to compare voxels to each other, we normalized these frequency tuning curves to make them independent of the total variance explained. Thus, the tuning function was defined

as the square of ρ for each specific frequency divided by the sum of the square of these ρ's over all frequencies **Figure 5.4A** (top 3 plots show examples for individual voxels). Thus, two voxels that show the same tuning profile but different maximum explained variance would have the same tuning curve. A tuning value of 0 for a given spatial frequency means that the associated feature maps explained none of the variance in activation across stimuli; a tuning value of 1 means it uniquely explained all of the variance.

Averaging the tuning distributions of all voxels within a ROI produces a tuning distribution at the level of ROI. Averages (dots in **Figure 5.4A**, bottom and **Figure 5.4**, **Figure 5.6B-C**) and error estimates were obtained by sampling with replacement 100 times the 16 validation subsets and associated encoding models for each voxel and then averaging across all iterations and voxels in a single ROI.

Consistent with previous studies (Henriksson et al. 2008), ROI-level tuning curves were found to empirically obey a log-Gaussian relationship. We thus performed nonlinear regressions to fit curves of this form to each tuning curve (curves **Figure 5.4A-C**). This fit was used to estimate the peak frequency values of the tuning curves (**Figure 5.4D**) and its shift (**Figure 5.4E**). The error estimate on the difference in peak frequency between imagery and vision tuning curves takes into account the uncertainty in the fitting procedure as well as the uncertainty in the ROI tuning points.

To obtain significance estimates for **Figure 5.4E** we tested the hypothesis of a non-zero difference between imagery and visual peak spatial frequencies against the null

hypothesis of no difference between peak frequencies. To construct the null distribution of peak frequency shift, we randomly shuffled the "vision" or "imagery" designation of voxel-wise tuning and calculated the mean frequency shift 1000 times. The region outside the red shaded area in **Figure 5.4E** indicates values with $p < 0.01$ for combined subject data. For each subject and point individually, the points with $p < 0.01$ are indicated by a black asterisk.

Stimulus Identification

An important measure of the validity of an encoding model is how well it can discriminate target stimuli that correspond to the measured activity from other "lure" stimuli, or in other words, how well it can decode imagined images (Naselaris et al. 2014). Here we used pairwise "hits" as a measure of identification accuracy. A "hit" occurs when the measured voxel activity pattern for the cued target is more correlated with the predicted activity pattern for that same target than the predicted activity pattern for a non-cued lure target. Note that any visual stimuli (not just the ones used in the present study) could be used to build the lure set of images, just so long as they could be fed to the encoding models to produce a prediction. However, for simplicity we selected our lure images from stimuli seen in the experiment. Identification accuracy for a given target stimulus is the percentage of hits accumulated across all lure images.

We performed two distinct types of identification. *Position* identification was used to determine if the encoding models successfully captured the way that object position was encoded in population activity (i.e. identify which of the 8 positions an object was imagined). Similarly, *object* identification was used to determine if the encoding models

successfully captured the way that specific objects, independent of position, were encoded in population activity (i.e. identify which one of the 64 different objects was imagined). For both types of identification, the cue was not included as part of either the target or lure stimuli. Thus, model predictions did not include any information about the cue (and I again emphasize that cue-responsive voxels were not included in this analysis, see section on Voxel Selection).

Both identification of locations and of objects is performed in two parts. First, half of the activation samples (256) for each voxel were randomly selected (but balanced such that all locations or objects were represented in each set). This half of the samples was used to estimate a cross-validated prediction accuracy score (Pearson correlation coefficient) for each voxel. These scores were used to rank-order all voxels from low to high accuracy and subsequently split them into groups of 500. The second part involved using each group of 500 voxels to calculate identification accuracy on the remaining half of the activation samples (256). In order to factor out the contribution of position during object identification and vice-versa, the activity patterns of samples that corresponded to the same identification target (a position or an object) were concatenated across either all 64 objects (for position identification) or all 8 positions (for object identification). This produced 8 series of concatenated voxel predictions and measured values for location identification, and 64 for object identification. We then evaluated the correlation matrix between all prediction series and measured series (see **Figure 3.6** for an example of an object identification matrix). The percentage of identification "hits" is then simply the

fraction of entries for which the diagonal elements have a greater value than the other entries in the row corresponding to the "lure" predictions.

To produce the curves shown in **Figure 4.2**, the previous procedure is repeated 100 times for overlapping brackets of decreasing voxel validation accuracy (i.e., the top bracket contains the 500 voxels with the most accurate encoding model predictions, the second bracket contains the voxels with the 251^{st} - 750^{th} most accurate, etc.). Values on the x-axes of plots in **Figure 4.2** give the largest (over the 100 repeats) of the smallest validation accuracy within each bracket. The standard error captures the variation of the identification "hits" percentage within each bracket.

To estimate the level of identification due to chance, the real identities of the locations (objects) underwent 5 shuffling per each of the 100 repeats discussed above and a common histogram was built from these 500 values for each bracket. Chance level for hits, the center of the null distribution, is always 50%. The region outside the grey shading near the bottom of plots in **Figure 4.2** correspond to identification score with significance level $p < 0.01$.

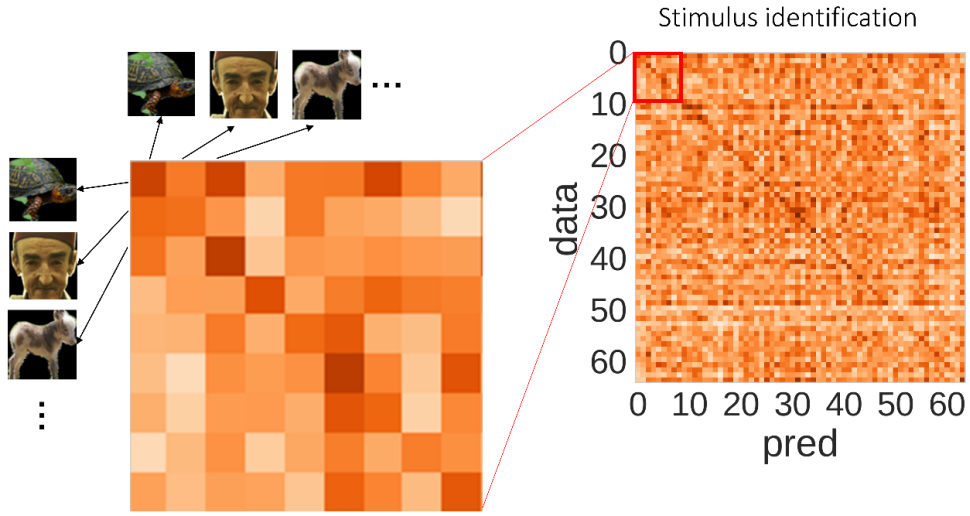

Figure 3.6 *Object Identification Matrix.* An example object identification matrix for one subject where each entry corresponds to the degree of correlation between the voxel activations predicted by the imagery encoding model for a single object picture (concatenated across all 8 positions) and the measured voxel activations in response to imagining the same object picture (darker orange is a stronger correlation). A "hit" is then counted every time the diagonal entry is higher than another "lure" entry along the same row.

Chapter 4 : Validation of the Imagery Encoding Model

Specific Aim 1: *Build voxel-wise imagery encoding models.* Hypothesis: imagery encoding models, like visual encoding models, can be successfully measured from fMRI data, used to predict brain activity to new imagined stimuli, and used to decode the position and content of the imagined stimuli.

Overview and Rationale

In the context of visual neuroscience, encoding models are models that attempt to capture the transformations that turn seen pictures into evoked brain activity. In human neuroimaging, *voxel-wise* encoding models can be built by examining the relationship between changes in visual features seen by a subject and systematic changes in individual voxel activity. Importantly, these encoding models make *predictions* about how a voxel will respond to a given input/stimuli and the model is in fact trained by tweaking its parameters so that it makes predictions that are closer and closer to the observed activity of that voxel. Once the appropriate parameters (those that lead to good predictions) are learned, those parameters can be used to infer "what" or "where" in the visual world a given voxel is tuned to. In other words, voxel-wise encoding models can be used to estimate the way in which visual features are encoded in individual voxels during a certain task. For the task of vision, such models are now routinely estimated (Kay et al. 2008; Naselaris et al. 2009; St-Yves & Naselaris 2017; Kay, Winawer, et al. 2013). For example, a relatively simple and popular encoding model is the population receptive field (pRF) model (Dumoulin & Wandell 2008) which estimates receptive fields (RFs), areas of the visual field that, when a stimulus is presented within it, evokes increased activity in

a neural unit such as a neuron. The pRF model estimates RFs for populations of neurons (i.e. a voxel) by presenting subjects with a high contrast stimuli that sweeps across the visual field as a bar or expands as a ring. The parameters of a given voxel's receptive field, specifically its size and location, are then adjusted until its prediction of a voxel's activity evoked by the stimuli closely matches that of the measured voxel activity evoked by that same stimuli.

The relationship that we have derived between vision and imagery (that imagery activity can also be expressed as a function of s) suggests that we should be able to recover imagery encoding models in the same manner as with visual encoding models only with a slightly different conditioning. That is, instead of viewing s, imagine s (see Equation 6). These encoding models could in turn be used to probe the hypothesized differences between vision and imagery. In the following chapter I give a brief overview of the experimental design that we implemented to vary visual features in *imagined visual space* while measuring the corresponding changes in fMRI voxel-wise activation patterns, and how this data was used to build an imagery encoding model (iEM) independent of a visual task

Methods

We measured whole-brain fMRI BOLD activity as three participants viewed and then imagined previously memorized object pictures in 1 of 8 different positions within the visual field. Each picture was associated with a six-letter cue and subjects familiarized themselves with picture-cue pairs prior to scanning. During vision runs, subjects were presented with both the cue (on which they fixated) and the picture in one of the positions while they passively viewed. During imagery runs, subjects were presented with the cue alone on which they fixated while imagining the matching picture. The color of the cue corresponded to the color of 1 of 8 brackets (each framing a portion of the stimulus field) and indicated the location in which the subject was to imagine the picture. For example, the cue "firtrk" written in blue means imagine the fire truck in the bottom left corner within the blue bracket. It is important to note that the colored brackets remained constant throughout all runs and the color-coded fixation-cues appeared in both conditions so that the only difference between vision and imagery conditions was the complete absence of the object picture during the imagery runs. The only source of variance during the imagery runs would therefore be the small changes in color and composition of the 6-letter cue at the center (overall size of the cue remained constant) and importantly, the imagined image.

To characterize the tuning properties of voxel activity during imagery, we used an encoding model estimation procedure developed in house: the feature weighted receptive field (fwRF) model (St-Yves & Naselaris 2017). The fwRF model is capable of describing not only the location and extent of receptive fields but also tuning to any

feature that can be extracted from the stimuli, such as spatial frequency. This model has been shown to recover receptive field properties and tuning functions consistent with known organizational principles of the visual cortex. The fwRF was used to estimate voxel-wise visual encoding models from the visual runs and imagery encoding models from the imagery runs. Encoding models specified tuning to spatial frequency and a receptive field location and size for each voxel. Each voxel therefore had two independent models, a visual encoding model and an imagery encoding model. Any voxel sensitive to the cues was discarded. All experimental runs performed with subjects 1–3 were also performed on a 4th control subject (S4) where eye tracking was measured (using a SR Research EyeLink 1000 eye-tracker) to confirm that any results obtained in relation to the stimuli (whose location was intentionally manipulated) could not be accounted for by eye movement. For more detailed information on the experimental design, data acquisition and processing as well the specifics of the fwRF model training see Chapter 3 on Experimental Methods.

Results

Analysis of Encoding Model Performance

A first and crucial test of the imagery encoding model (iEM) is to show that it can successfully explain variance in signal across mental images that were not used to train the model. To do this we performed a k-fold cross-validation analysis (see Methods for details). Briefly, the trained imagery models were used to generate predictions of voxel activity in response to held-out imagined pictures sets. The validation was then scored as the Pearson correlation coefficient (CC) between the predicted and measured activities

during imagery for each voxel. As a reference point, this was also done for the visual encoding models (vEMs) using held-out data from the vision runs. Therefore, each voxel had a validation score for the iEM and a validation score for the vEM. **Figure 4.1** shows the cross-validation prediction accuracy for each voxel from the vEM (top row) and iEM (middle row) mapped onto a flattened cortical surface of each subject, while the bottom most row shows the distribution of iEM scores (orange) and vEM scores (blue) for all voxels as well as their joint distribution (green). All voxels falling above the horizontal grey line reached the significance threshold set on prediction accuracy (Pearson correlation coefficient $\geq .16$, $p < 0.01$ see methods for how this was calculated). The iEM was able to predict activation in response to imagined pictures well above significance for many voxels in all visual cortical areas considered here (see **Error! Reference source not found.** for count of significant voxels by visual area for all subjects).

To further establish the validity of the iEM we performed model-based identification. That is, how well the model can pick out the cued imagined stimulus from other "lure" stimuli using predictions of activity patterns across voxels. We distinguish between two measures of identification performance: position identification (i.e. which of the 8 positions the object was shown) and object identification (i.e. which one of the 64 different objects was shown, regardless of its position). Most importantly, *object* identification must rely on some feature(s) of the object present in the stimulus, not just the area of the visual field that it happens to occupy, and therefore its success can rule out

certain confounds (such as spatial attention). In both cases, we used the pairwise accuracy ("hits") to represent identification accuracy.

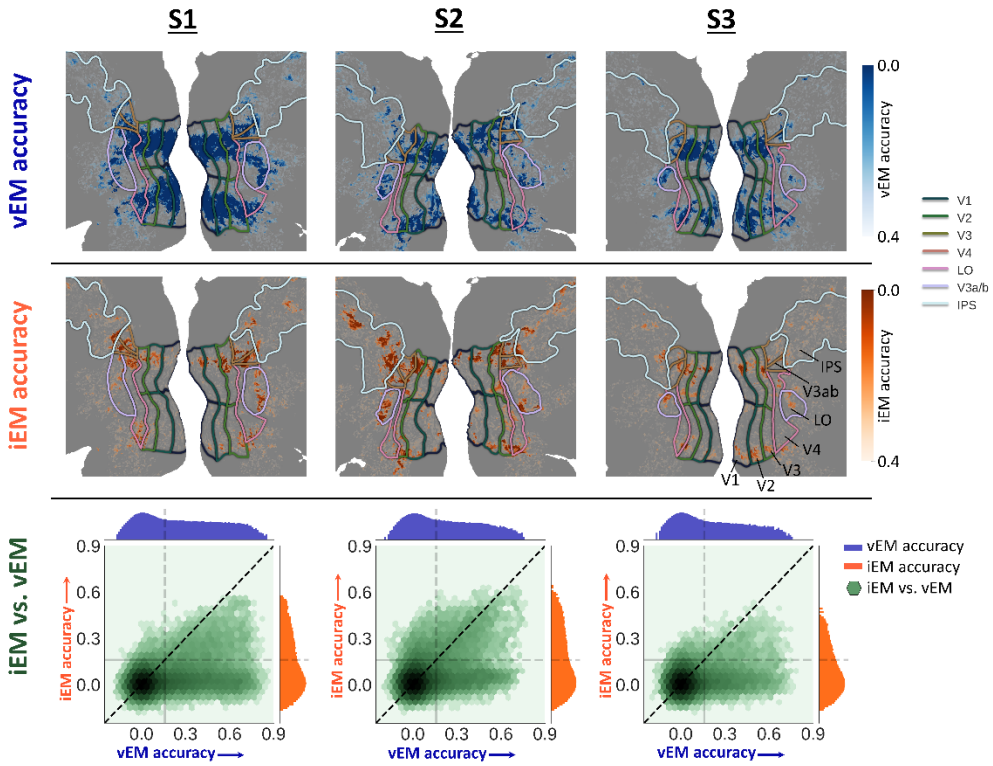

Figure 4.1 *Cross validation accuracy of encoding models.* Prediction accuracy (colorbar) of the visual encoding model (vEM; top row) and imagery encoding model (iEM; middle row) mapped on the flattened cortical surface for each subject. Bottom: joint histogram (green) and marginal histogram of prediction accuracy for imagery (orange) and visual (blue) encoding models across all voxels for subjects 1-3. The iEM makes accurate predictions of imagery activity (Pearson correlation ≥ 0.16, $p < .01$; dashed grey lines) in all subjects.

Briefly, identification was assessed in groups of 500 voxels at a time, and for every cued target (an imagined object or position) a pattern of evoked activity was measured in the brain across all the voxels in that group. A hit was counted whenever the *measured* group pattern for a cued target was more correlated with the *predicted* group pattern for that same cued target than for some other non-cued "lure" target (see **Figure 3.6** for example matrix of correlations). Predictions of activity patterns made by the imagery encoding models were accurate enough to identify position of (**Figure 4.2**, top) and object in (**Figure 4.2**, bottom) the imagined stimuli. This demonstrates that subjects were imagining the cued objects as instructed. Moreover, groups that contained voxels with higher prediction accuracies performed better identification. This suggests that the success of the model depends, to a degree, on the object-picture contained in the stimuli. Accurate identification of imagined objects would not be possible if variation in spatial attention, eye position or visual cues were the sole determinants of prediction accuracy.

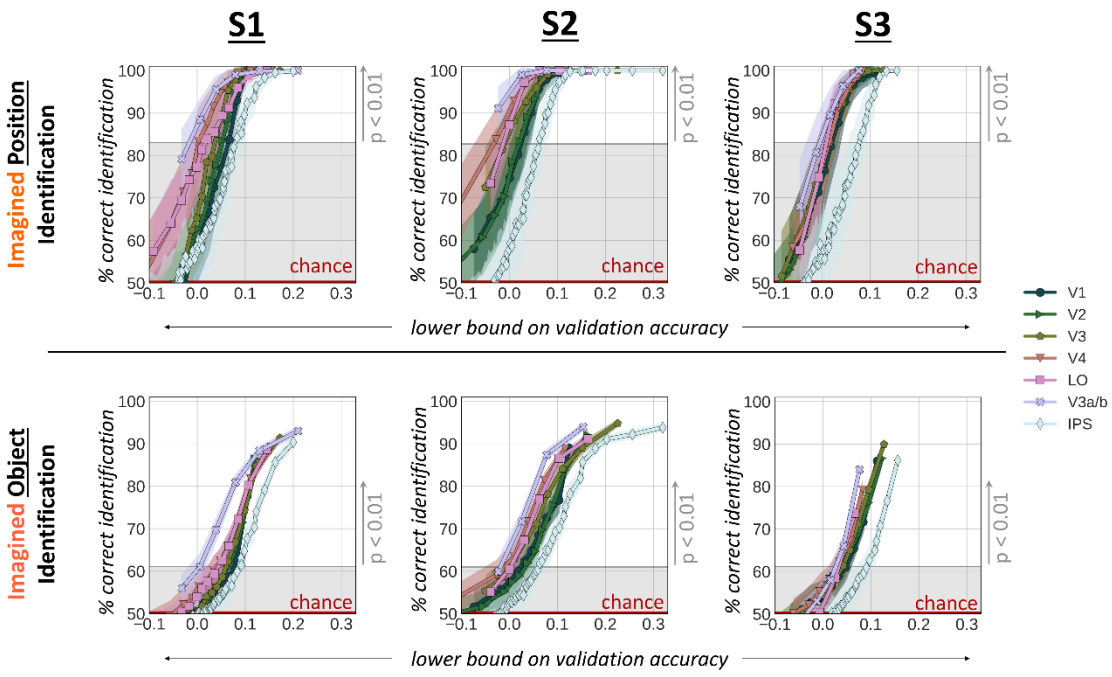

Figure 4.2 *Identification of imagined stimuli***.** Model performance in identifying the correct position (top) and object picture (bottom) of the imagined stimuli for each ROI in each subject. Curves show percentage of correct pairwise identification (colored shading indicates $\pm\, SE$; gray shading indicates statistical significance threshold of $p < .01$ (permutation test) for subpopulations of 500 voxels in visual area. Ordering along x-axis is by lowest prediction accuracy of all voxels in each subpopulation.

Analysis of the Learned Encoding Model Parameters

The goal of aim 2 is to measure changes in encoded features across the visual cortex from vision to imagery. In order for this comparison to have meaning, we must first make sure that our model is assigning receptive field and tuning parameters that reflect known functional structure of the visual cortex during vision. Cortical maps of receptive field size, receptive field location and peak spatial frequency derived from visual encoding models (**Figure 4.3**, left) are consistent with maps observed in many previous studies, e.g., (Dumoulin & Wandell 2008; Kay et al. 2015; St-Yves & Naselaris 2017; Hansen et al. 2007)**.** Namely, receptive field size increased and spatial frequency preference decreased with distance from foveal representations, and reversals in receptive field visual angle occurred over the boundaries between ROIs (note that ROI boundaries were drawn using a separate set of standard retinotopy mapping experiments). Interestingly, for every subject, the imagery encoding models also exhibit reversals at the boundaries of visual ROIs (**Figure 4.3**, top right) that are consistent with visual organization. Plots of visual receptive field locations show expected relationships between visual field quadrants and ventral/dorsal, left and right cortical hemispheres (**Figure 4.4**). Additionally, visual encoding models reproduce expected size-eccentricity relationships (**Figure 4.5**).

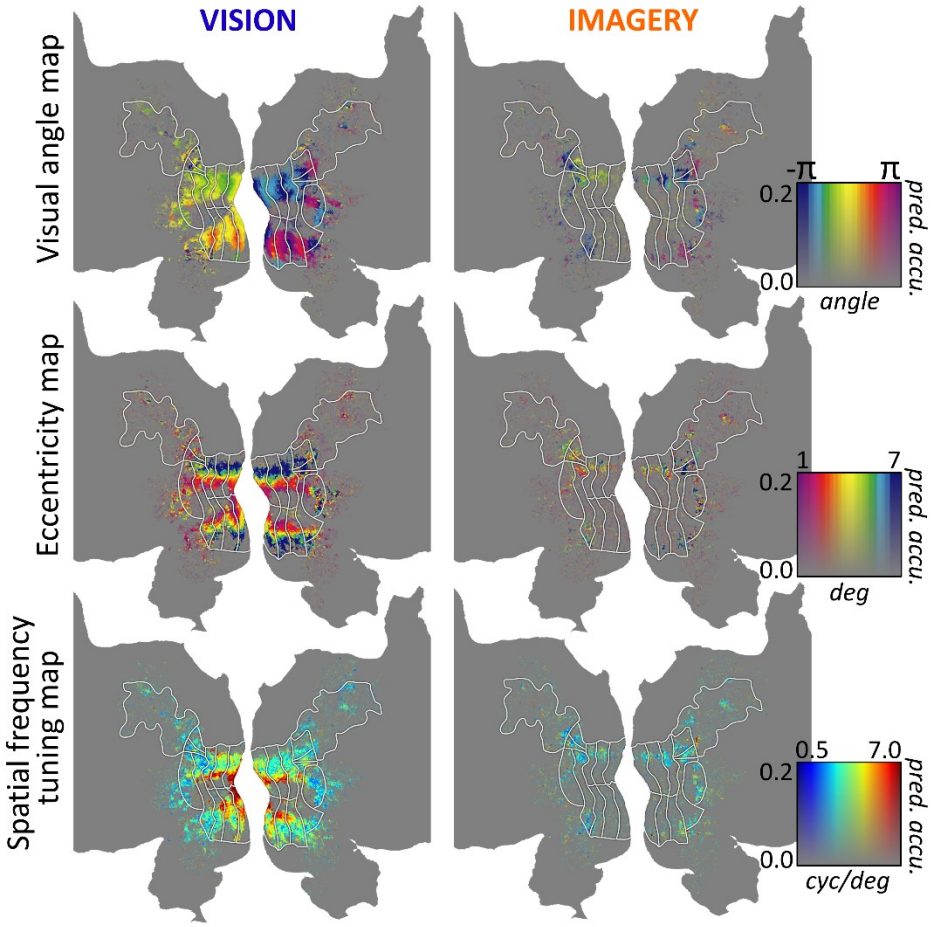

Figure 4.3 *Anatomical layout of encoding model attributes.* Visual angle (top row), average eccentricity (middle row), and spatial frequency tuning (bottom row) during vision (first column) and imagery (second column) shown on a flatted cortical surface for subject 1, displaying the retinotopic organization recovered by the encoding models.

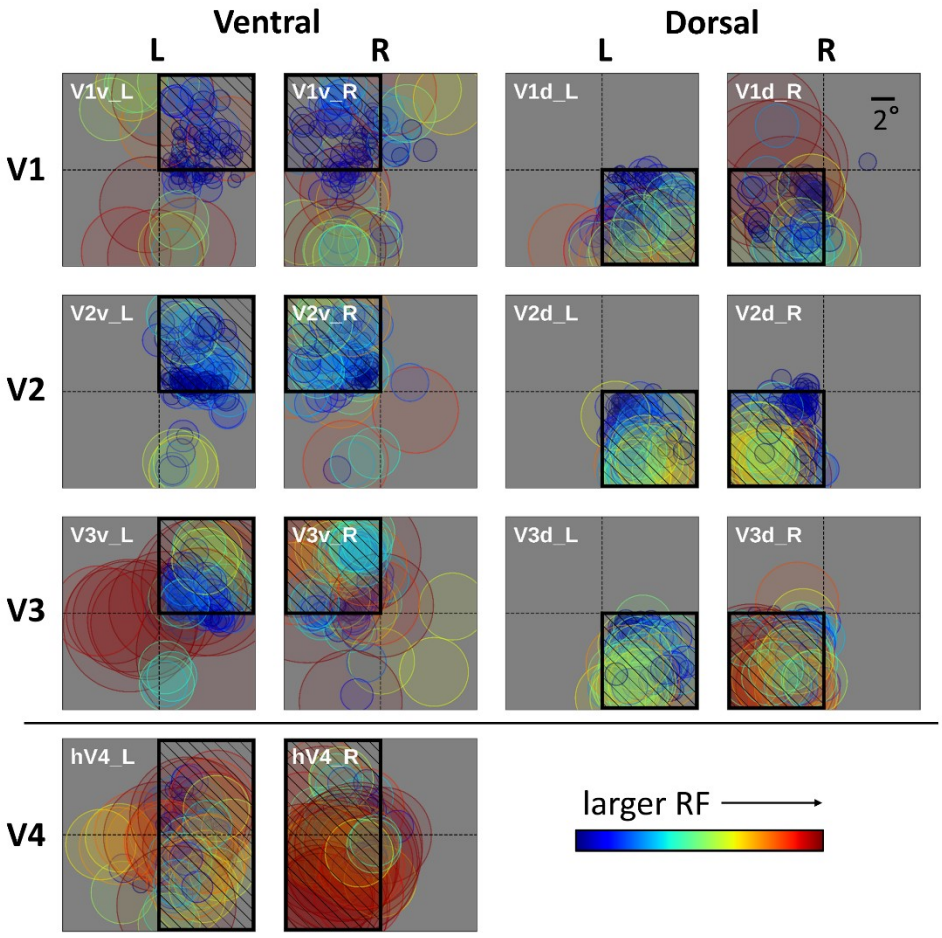

Figure 4.4 *Visual receptive fields***.** Circle plots showing the average raw receptive fields (locations and relative sizes) corresponding to a subset of voxels from each area during vision. Each ROI (row) is partitioned by hemisphere ("L" = left and "R" = right), and/or dorsal ("d") and ventral ("v") position in order to demonstrate specificity of model receptive field properties by quadrant of visual field. The dashed lines delineate the four quadrants of the visual field and the black hatched areas fill in the quadrant or side one would expect to see a concentration of receptive field locations (given known retinotopic organization of the human visual cortex).

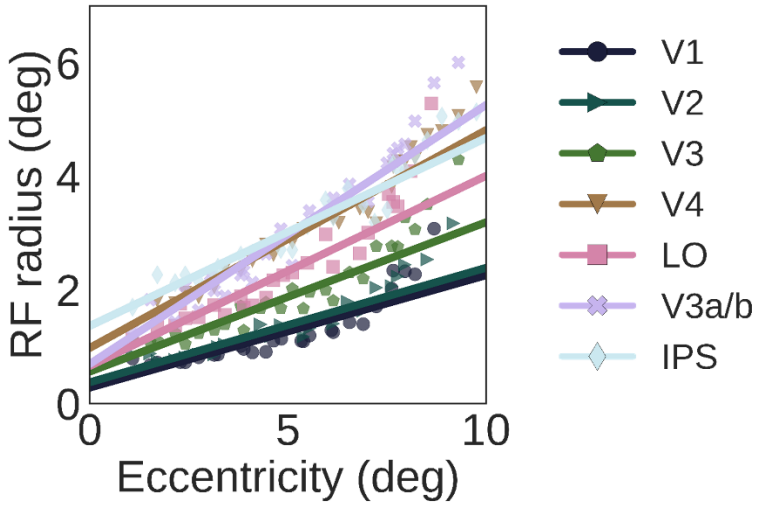

Figure 4.5 *Size-eccentricity relationships.* The assigned visual receptive field size as a function of the assigned receptive field eccentricity for voxels from each ROI. In line with known organizational properties of the human visual cortex, this relationship becomes steeper with ascent of the visual hierarchy.

Control for Potential Eye-Movement Confounds

As is shown above, the prediction of the imagery encoding models were accurate enough to identify imagined object pictures independently of their location (**Figure 4.2**). This result in itself can rule out any confound due to subtle eye movements in the direction of the imagined scenes. However, to demonstrate this directly, we ran our experiment with a fourth control subject while tracking the subject's eye movements. We estimated and then validated imagery encoding models for this subject, obtaining results comparable to those obtained for subjects 1-3 (**Figure 4.6**). The following steps were taken to show that the subject did not have biased eye movement which might explain these imagery encoding model results.

First, to confirm that the eye-tracker read-out was showing the correct relative location of the eyes fixation within the visual field, we ran an extra test run (in addition to standard calibration that was performed at the start of each run) in which a target dot appeared in 1 of 9 known positions spanning the area in which the experimental pictures were shown. The subject was instructed to fixate on the target as it appeared. **Figure 4.7A** shows that the location of measured eye fixations (colored points) overlapped with the location of the target (marked with an X). Eye fixations were then measured during each vision and imagery run. As demonstrated by the example imagery run in **Figure 4.7B**, the measured fixations were not systematically shifted relative to the different locations that the subject was instructed to imagine the object pictures.

Next, to account for any *overt* eye movements that were not observable but still systematically varied with the stimuli, we created a simple linear model that attempted to

predict brain activity using eye fixations as input. For each run (n = 21) the simple linear model was trained on three-quarters of the data, where the target was the GLM beta weights assigned to a given condition (a picture and location) and the inputs were the x and y of the eye fixations during that condition. The resulting model weights were used to predict the beta values of the remaining left-out conditions (while taking care that this validation group had 2 of each position). These predictions were compared to the measured betas from the brain activity, giving a correlation coefficient (ρ) for each voxel. The flatmap in **Figure 4.7C** shows the average run performance score (ρ) for the eye-fixation model for each voxel. The scale is the same as those shown in main in **Figure 4.1** and **Figure 4.6**, demonstrating that there are no areas in which eye fixations have prediction power. Finally, to ensure that the above eye-fixation model failed because there were no systematic biases in eye fixations, rather than failing due some other inherent issue of the model, we created "synthetic" eye fixations that varied in sync with changes in the position of the stimulus. When given this fabricated and systematically-biased data, the model performed quite well (**Figure 4.7D**) in the visual cortex, showing that results in **Figure 4.7C** were due to a lack of overt biases in eye movement that might account for our main results. This further highlights the importance of eye tracking and other controls (i.e. the image identification) for potential confounds.

Taken together, these results show that the subject fixated at the center of the visual field as instructed, no matter the position of the imagined stimulus, that there was no evident systematic bias in eye fixations related to position of stimulus, and that even with

eye movements controlled for, we still recovered the same imagery encoding model validation results that were observed in subjects 1–3 (**Figure 4.6**).

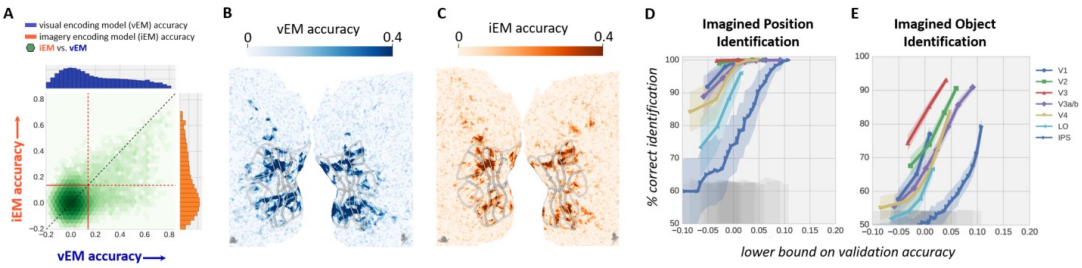

Figure 4.6 *Cross validation accuracy of encoding models and stimulus identification for control subject with eye-tracking.* Results are displayed in a similar fashion as **Figure 4.1** and **Figure 4.2** to demonstrate that data acquired at 3T with eye-tracking replicates the finding of the three subjects in the main experiment. A) Validation accuracy during imagery as a function of the validation accuracy during vision for each voxel (hexbinned). Marginal plots show the individual distributions of model validation for vision (blue) and imagery (orange). The red lines indicate the validation accuracy threshold, where the lower-right quadrant corresponds to voxels that are well predicted exclusively during vision while the upper-left quadrant correspond to voxels that are well predicted exclusively during imagery. The upper-right quadrant corresponds to voxels predicted in both modalities. Visual B) and imagery C) prediction accuracy plotted on the flattened cortical surface. Darker colors indicate higher prediction accuracy. D) Model-based identification of imagined stimulus position accuracy from a subpopulation of 500 voxels within each visual area plotted against the lowest prediction accuracy within that subpopulation of voxels. E) Same as in (D) but with identification of imagined stimulus picture content. In both cases, the gray shaded area represents one standard deviation of the distribution of identification accuracy due to chance. The colored-shaded area around each curve correspond to one standard deviation of the accuracy estimate.

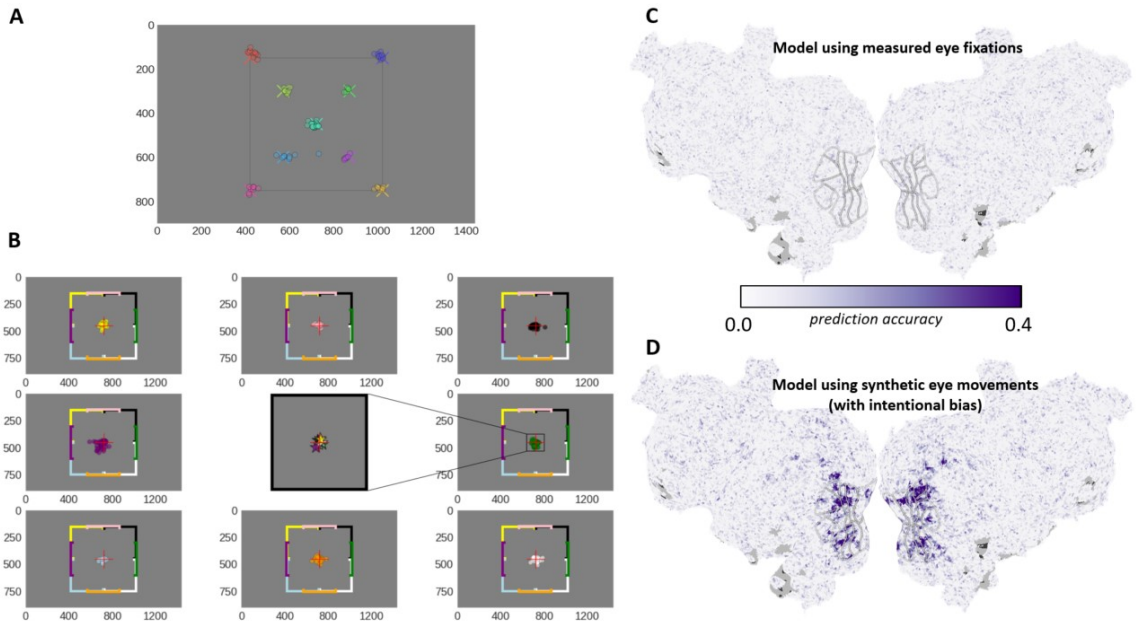

Figure 4.7 *Eye-tracking control results.* **A)** Eye-tracking calibration results demonstrating that the eye-tracker was accurately tracking eye fixations. Xs mark the positions within the visual field the target appeared (each location in a different color) during an eye-tracking calibration test. The colored circles indicate the location of the subject's fixation as measured by the eye-tracker, each colored to match the X marking the position of the target during that given fixation. **B)** Example eye fixations during the experiment demonstrating that the subject successfully fixated at center with no observable biases toward the location of the imagined picture. Plots show all fixations for all conditions in a sample imagery run separated by location of stimulus. The colored brackets are the same as was present for all imagery and vision runs. The color of the fixation points (as well as the relative location of the plot in which it appears) correspond to the position in which the picture occurred during that fixation. Center plot shows the center of mass of each group of fixations (color matches those in the other 8 plots) plotted together and zoomed in to 160x160 pixels to show their relative location in respect to each of the other positions. **C)** Flatmap showing the prediction performance of a linear model attempting to use eye fixations to predict brain responses to different stimulus positions. Scale is the same as those shown in **Figure 4.1** and **Figure 4.6**, demonstrating that eye fixations had no prediction power. **D)** Results that would have been obtained had there been systematic biases in eye movement. Prediction accuracy results when the eye-fixation model was given synthetic eye movements that varied in sync with the cued position of the imagined object picture. In this case the model performed quite well, ensuring that there was not an issue with the structure of the model itself, but that a total absence of systematic eye movements was underlying the lack of prediction power in (C).

Interpretation and Discussion

The iEM is the first instance of a feature encoding model estimated directly from mental imagery: previous studies examining encoded features in imagery used models trained on visual data and only tested on imagery (Naselaris et al. 2014; Senden et al. 2019). The iEM is built solely from fMRI data collected during mental imagery and is entirely independent from visual encoding model that we have built here for comparison. Our experimental design allowed us to remove virtually all retinal variance so that we could isolate the neural responses evoked from changes happening in the imagined visual space. Even with the only source of variance coming from internally generated imagery, the model was able to learn how to accurately predict voxel activity and to decode the content and location of new imagined stimuli in a manner similar to the established visual encoding model. We have thus demonstrated the feasibility and utility of the iEM as a novel tool for investigating mental images. These findings license us to use the iEM to directly infer the features that have been encoded during imagery across the visual hierarchy. The follow chapter explores the parameters of the imagery and visual encoding models and tests for differences in the encoding of imagined and seen stimuli.

Chapter 5 : Signatures of Inference in a Generative Model: Shifts in Properties from Vision to Imagery

Specific Aim 2: *Determine if signatures of inference in a generative model can be observed during mental imagery in the human brain.* Hypothesis: (1) tuning to imagined features in lower visual areas will more closely resemble tuning to seen features in higher areas and (2) a gradient in this distortional effect that increases with hierarchical distance below the reactivated area, demonstrating key signatures of inference that follow from our formulation of mental imagery as inference in an internal generative model.

Overview and Rationale

In Chapters 2 and 3, I laid out the theoretical and technical tools needed to build encoding models, and in Chapter 4 I established that such models can be constructed from brain data during imagery alone and that these models contain meaningful information about the features encoded during mental imagery. In this chapter, I describe how we used these tools to explore how vision and imagery compare and how this varies, if at all, across visual areas.

Following the relationships described in Chapter 2, if the HGM is a good model for both forms of visual experiences then we expect to see a gradient of increasing distortions moving down the levels of the visual hierarchy whereby receptive field and tuning properties of imagery shift away from vision properties for that level, and towards the vision properties of the source (clamped) area. Also note that the echo effect should only cause a distortion in all areas of the hierarchy *below* the clamped stage. Because of the assumed lossless reactivation of the clamped stage and the structure of the hierarchy

where each stage only "sees" the activity of the stage directly above or below it, inference will proceed above the clamped level exactly as it does during vision. Therefore, all areas above the clamped stage will converge to an activation state during imagery that is indistinguishable from vision, and increasing distortional effects will be seen below the clamped level. If such patterns between vision and imagery do exist we should be able to detect them using encoding models built and tested in Chapter 4.

The specific features one would expect to find encoded druing mental imagery then follow from our hypothesis combined with the known organizational principles of encoded features during vision across the visual cortex. During vision, the types of features encoded in higher-level areas in the human visual hierarchy tend to be more abstract and of lower resolution relative to lower-level visual areas (**Figure 5.1**, left). For example, high-level areas respond preferentially to low spatial frequencies and are tuned to large portions of the visual cortex (i.e. they have large *receptive field*s), while low-level areas respond preferentially to high spatial frequencies and have small precise receptive fields for working out high resolution details (Henriksson et al. 2008; Dumoulin & Wandell 2008; Kay, Winawer, et al. 2013; Grill-Spector et al. 2018).

Therefore, we expect to find increasing divergence in such properties moving down the visual hierarchy whereby imagery receptive fields become relatively larger and more foveal and imagery spatial frequency preferences become relatively decreased (**Figure 5.1**, right).

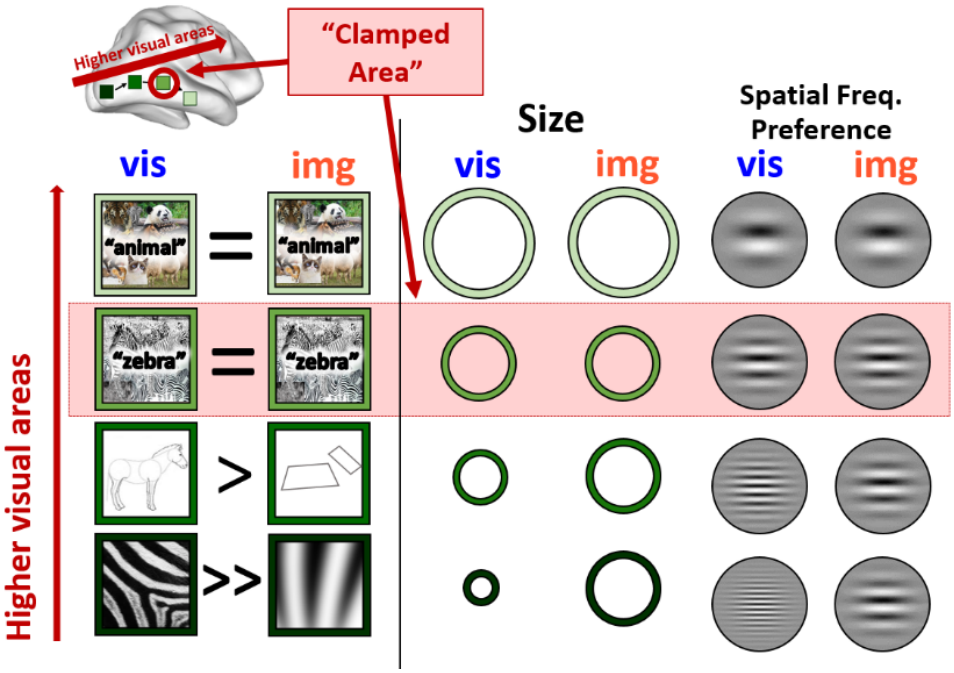

Figure 5.1 *Hypothesized changes in receptive field and tuning properties*. Schematic illustrating the change in resolution and encoded features (such as receptive field size and spatial frequency preference) during vision (blue) with ascension of the visual hierarchy. We hypothesize that at and below the clamped area, the features encoded during imagery (orange) will more closely resemble those of the clamped area during vision, resulting in a loss of resolution in the lower, more detail-oriented layers.

Methods

Detailed descriptions of how the imagery and visual encoding models were built from the fMRI BOLD signal can be found in Chapter 2. Both encoding models specified a separate receptive field location and size as well as spatial frequency tuning for each voxel (see **Figure 4.3** for an example of the retinotopic organization of these estimates).

In order to assess how the encoding model attributes changed from vision to imagery across the visual hierarchy we defined regions of interest (ROIs) within the visual cortex for each subject. This was done using separate retinotopic mapping runs and standard probabilistic topographic maps. Differences between vision and imagery in terms of model accuracy, SNR, spatial frequency tuning, receptive field size and location were then calculated for each ROI in each subject.

Results

Imagery Encoding Model Prediction Accuracy and Signal-to-Noise Exhibit Graded Attenuation Across Hierarchical Levels.

If during imagery an activity pattern in one visual area is clamped to an expected visual activity pattern, we should expect prediction accuracy of imagery and visual encoding models to be close to parity in this area. This was true in intraparietal sulcus (IPS), a collection of visual areas at the highest level of processing considered here (**Figure 5.2A** and **Figure 5.3**). Relative prediction accuracy of the imagery encoding model decreased with descent toward primary visual cortex (V1). This gradient is highlighted by noting that the slope of purple line (the best linear fit of iEM to vEM) approaches the line of parity (iEM accuracy = vEM accuracy) from low to high areas.

The distance from parity (slope of fitted line – 1; less negative numbers indicating stronger similarity) is plotted for each ROI and each subject in **Figure 5.2B**. The gradient in relative encoding model prediction accuracy is most likely due to the matched gradient in relative signal-to-noise (SNR; **Figure 5.2C**). The attenuation of SNR during imagery in the brain tracked an attenuation in signal amplitude (**Figure 5.2D**). Interestingly, this finding is in line with the hierarchical formulation of a generative model, as a loss of signal would be expected with each transformation taking the activation from the clamped level to lower levels. Noise was uniformly reduced during imagery at all processing levels and for all subjects (**Figure 5.2E**), a result that could be a consequence of clamping an additional stage during imagery which would effectively reduce the number of random variables in the system and therefore reduce noise. This prediction accuracy and SNR gradient serves as the first indication that the relationship between imagery and vision varies systematically across the visual cortex in line with the echo hypothesis.

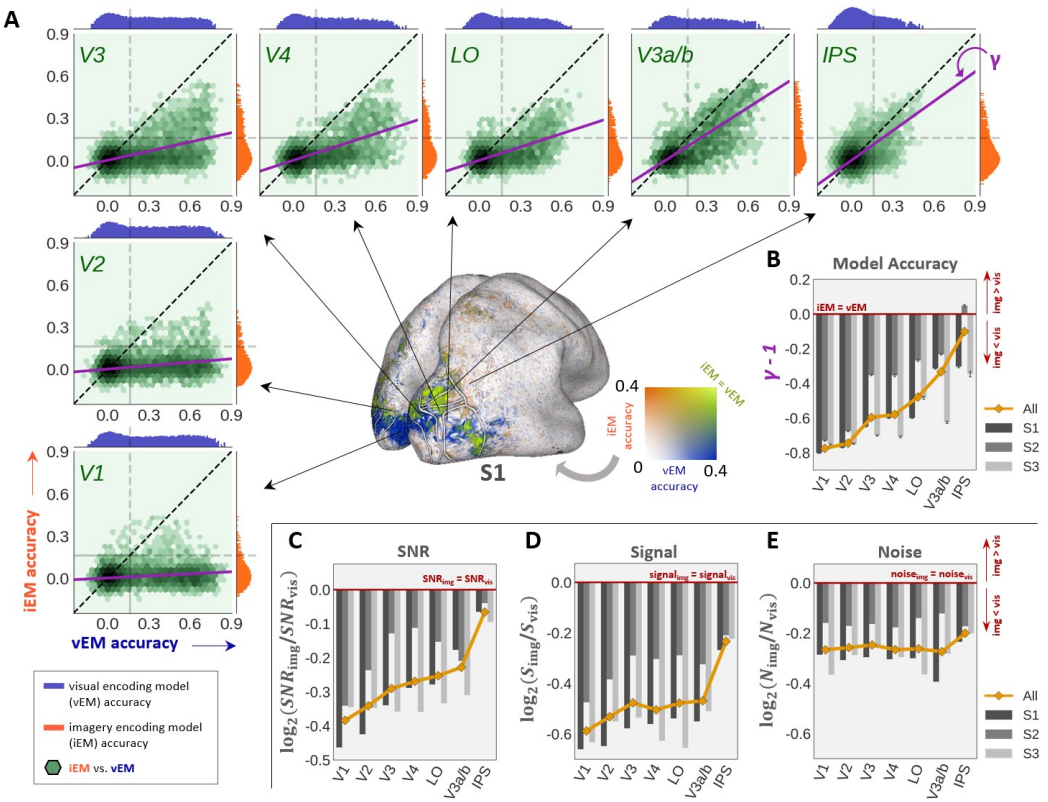

Figure 5.2 *Relative prediction accuracy of imagery encoding models (iEM) across visual areas.* (**A**) joint histogram (green) and marginal histogram of prediction accuracy for imagery (orange) and visual (blue) encoding models for indicated area (subject 1 only; ordering of visual areas follows (Markov et al. 2013). Purple line shows slope (γ) of best linear fit of iEM to vEM prediction accuracy. Inflated brain surface map shows relative prediction accuracy (2d colormap) of the iEM and vEM. (**B**) Difference from parity (γ -1) for each area. (**C**) Median signal-to-noise ratio (SNR) for imagery activity relative to visual activity. (**D**) Relative signal (*S*; i.e., activation amplitude) and (**E**) noise (*N*).

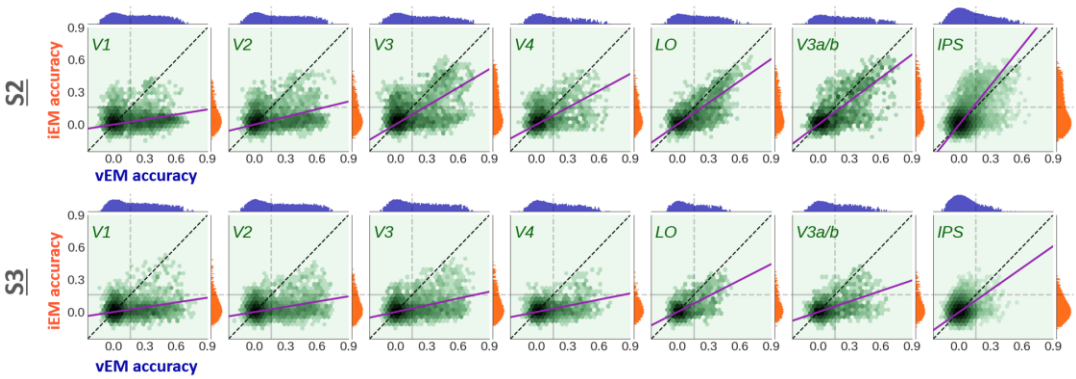

Figure 5.3 *Relative prediction Accuracy for subjects 2 and 3*. Format as in **Figure 5.2**.

Spatial Frequency Preference During Mental Imagery is Reduced Relative to Vision in Low-Level Visual Areas

The theory predicted that spatial frequency preference during imagery should decrease relative to visual spatial frequency preference with descent from the clamped level toward V1. Such decreases in spatial frequency preference during imagery relative to vision were in fact observed **(Figure 5.4**). Unlike encoding model prediction accuracy, loss of SNR in early visual areas cannot account for these effects. In other words, it is not the case that the model automatically assigns lower spatial frequency preferences to noisier voxels potentially leading to the observed shift in tuning for imagery. Rather than being somehow inherently linked to low frequencies, noisy voxels instead exhibit flat turning curves across all areas for both imagery and vision (dashed curves in **Figure 5.5**).

In order to demonstrate the generality of the tuning shift phenomenon, we considered two different "spatial scales": overall ROI population tuning shift and voxel-wise tuning shift. In the population shift, an overlapping but not necessarily identical group of voxels may be used (i.e. mean tuning of all the voxels in V1 that are well predicted during vision vs. mean tuning of all the voxels in V1 that are well predicted during imagery), while the voxel-wise shift can only be assessed in voxels that were well predicted by both vision and imagery (i.e. mean of shifts from vision to imagery within voxels). The two perspectives are indicated by the inset Venn diagrams in **Figure 5.4B-D** (individual shading of orange or blue circle indicate population tuning, and shading of the overlap in green indicate voxel-wise tuning) and can be linked to changes in activation amplitude during imagery in response to different spatial frequencies. If during imagery the amplitude decays uniformly across all neurons and features, it may be possible to see a

tuning shift at the level of ROI (already low activations during vision could drop low enough during imagery to have voxels removed from the imagery tuning average), but we *would not* expect to see any tuning shifts at the level of an individual voxel. Yet if the amplitude decay during imagery "spares" specific populations of neurons, we would expect to see tuning shifts at both the population level and the individual voxel level. As is shown in **Figure 5.4B** and **C** the spatial frequency tuning shift from vision to imagery is seen at both spatial scales. This is consistent with selective neuronal population changes in early visual areas leading to shifts toward the frequency tuning of the clamped area. All subsequent panels and analyses use the voxel-wise population.

Receptive Field Location and Size are Altered During Imagery Relative to Vision in Low-Level Visual Areas

Another predicted effect of the echo transformation is that imagery receptive fields should be increasingly dilated and displaced toward the fovea relative to visual receptive fields with descent toward V1. In V1, imagery receptive fields were larger (**Figure 5.6A,B**) and more foveal (**Figure 5.6C,D**) relative to vision for each subject. Consistent with our theory, the evidence for differences between imagery and visual receptive field attributes weakened with ascent toward high-level visual areas.

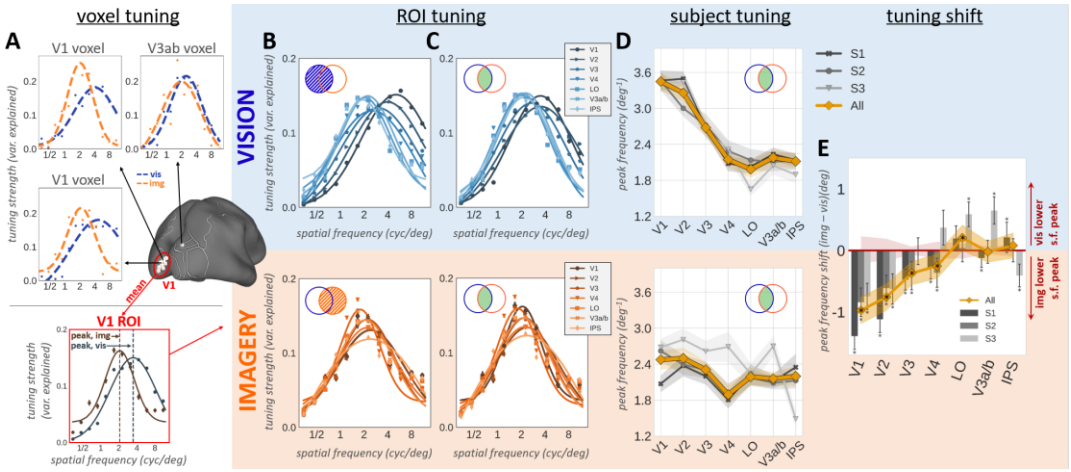

Figure 5.4 *Differences in spatial frequency tuning between vision and imagery*. Differences in spatial frequency tuning between vision and imagery. (**A**) Visual (blue) and imagery (orange) spatial frequency tuning curves for single voxels sampled from V1 and V3ab and population tuning curves for V1 (bottom). (**B**) Top: Population tuning curves during vision for all voxels in the indicated area that have an accurate vEM. Bottom: Population tuning curves during imagery for voxels that have an accurate iEM. Populations in top (blue circle in Venn diagram) and bottom (orange circle) plots are overlapping but not identical. (**C**) Population tuning curves for all voxels in the indicated area that have an accurate vEM and iEM. All subsequent panels use this population. (**D**) Peak spatial frequency of tuning curves in (C). (**E**) Difference between peak spatial frequency during imagery and vision.

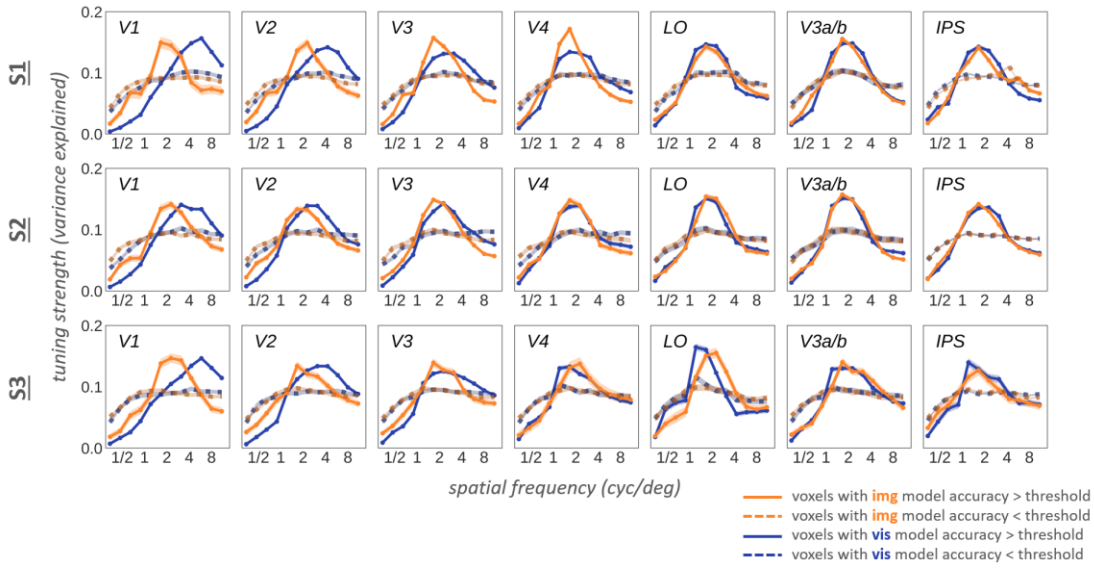

Figure 5.5 *Spatial frequency tuning curves for all ROIs and subjects.* Spatial frequency tuning curves for all ROIs and subjects. Plots show the average tuning curves in each ROI in S1-3 for vision (blue) and imagery (orange). Solid lines represent the tuning of all voxels with a prediction accuracy above a threshold while dotted lines represent the tuning curves of all voxels below threshold (i.e., voxels for which either the imagery or vision encoding model gave poor predictions). This demonstrates that our modeling procedure does not induce a bias toward low frequency preferences for voxels with substantial unexplained variance. Rather, for such voxels the modeling procedure produces flat spatial frequency "tuning".

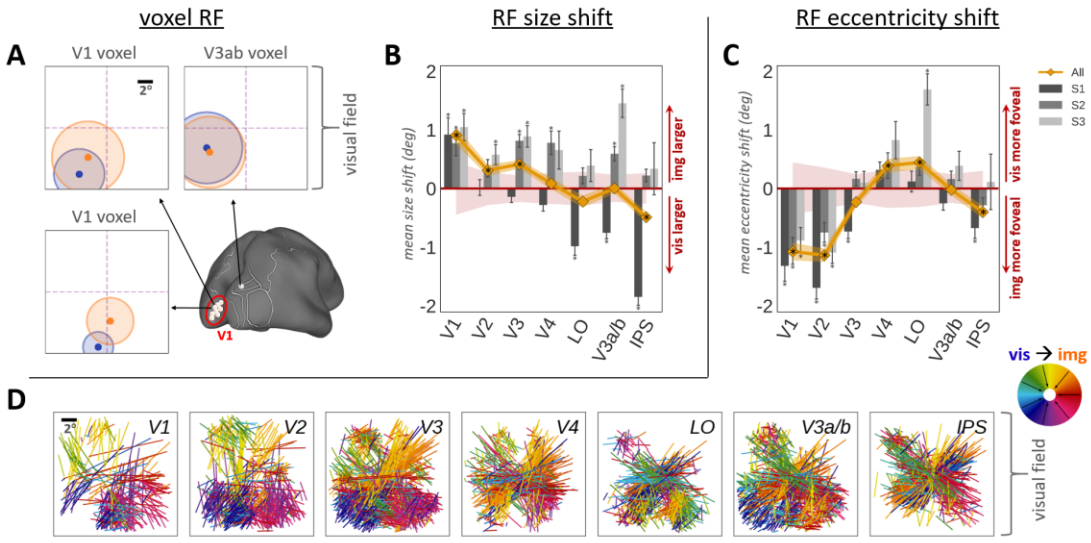

Figure 5.6 *Differences in receptive field location and size between vision and imagery.* (**A**) Example visual and imagery receptive fields (RF) for single voxels (**B**) Average signed change in RF size from vision to imagery. Positive (negative) values indicate dilation (shrinkage). (**C**) Average signed magnitude of shift in RF location from vision to imagery. Negative values indicate a shift toward fovea. (**D**) Orientation and magnitude (line segments) and direction (color-wheel at far right) of RF location shifts (same voxels as in (B) and (C)) from vision to imagery. The red shaded area in (B) and (C) indicates significance level $p < .01$ (permutation test) for combined subject data (yellow curve). In all panels asterix indicates significant difference from null value (red line, $p < .01$, permutation test; red shading indicates significance threshold for combined data); shading on curves indicates $\pm SE$.

Interpretation and Discussion

Summary of Results

We have shown, through building and leveraging imagery encoding models, that there is a gradient of distortion in imagery receptive field properties traveling down the visual processing stream. This distortion moves the representations encoded during imagery towards those seen in higher areas during vision. Specifically, imagery receptive fields are on average larger and more foveal, and tuning is shifted towards lower spatial frequencies relative to vision in lower-level areas. This is consistent with the echo effect that follows from our formulation of mental imagery as inference in an internal, hierarchical generative model. Furthermore, we have shown that, for the specific task performed by the subjects in our experiment, the level of clamping happens around area V4. The implications of this as well as other details and caveats of this study are discussed in detail in the following sections.

On the Level of Clamping and Use of Complex Stimuli

In this study we chose to use complex natural stimuli for two main reasons. First, such stimuli approximate what we see and imagine in real life (i.e., of the laboratory). Secondly, and more importantly, we chose to use complex stimuli because the purpose of our experiment was to test predictions about the differences between vision and imagery at all levels of the visual hierarchy, and complex natural stimuli are known to engage the visual system at all levels (Grill-Spector & Malach 2004; Einhäuser & König 2010; Çukur et al. 2013) in a subtly different manner than would the sum of their parts (Kayser et al. 2004; Carandini 2005; Snow et al. 2017). Complex stimuli thus satisfied a

fundamental requirement of our experimental design. Simple parametric stimuli (e.g., bars, gratings, dots) tend to robustly excite only low-level visual areas, and thus would have been a poor choice for our experiment. However, at first blush it might seem that using such complex stimuli may leave room for the subjects to "unsuccessfully imagine" the object (imagine a "zebra" but not imagine the details of the stripes) and might suggest that using simple stimuli (such as gratings, bars, dots vs. natural stimuli) would make it easier for subjects to "correctly imagine" the stimuli, subsequently closing the gap in the differences between vision and imagery that we found in lower areas. This idea however is based on a common but unfounded assumption and puts forth a scenario that is actually not inconsistent with our theory.

First, such an argument presupposes that we know what performing imagery perfectly would imply for the neural activity, suggesting that in order for one to successfully imagine something, they must imagine it just as it were during vision. This seems unlikely, given the phenomenological and measured differences between vision and imagery. Under the generative interpretation to "imagine poorly" simply means to clamp high. Any lack of detail in imagined images therefore does not indicate a failure of the subject to imagine correctly but rather reflects a general limit on our capacity to experience the details of a complex object we are holding in our mind's eye. In other words, our model describes why such a limit would exist. Such high-level clamping may help to explain why mental images lack the specificity of seen ones. High-level visual areas provide a poor substitute for the visual detail encoded in retinal activity during vision. Formulated as an echo transformation, inference conditioned on a high-level

representation of the stimulus will effectively low-pass filter the image representation. The strongest evidence for this effect comes from the reduced spatial frequency preference observed in V1 and V2 in all subjects during imagery. Therefore incidentally, our theory suggests a solution to the issue of *indeterminacy* discussed in the introduction: being able imagine a specific object without being able to imagine the details of that object is perfectly in line with clamping high. For the same reasons, our theory also explains why vision encoding models are somewhat successful at predicting and decoding imagery activity (Naselaris et al. 2014; Horikawa & Kamitani 2017), and characterizes further their intrinsic limitations.

Secondly, it is important to note that a decrease in the differences between vision and imagery with the use of simpler stimuli/tasks than the one used here does not contradict our account of imagery, and in fact our model would predict such a change given variations in the level of reactivation. This is because we do not claim, nor does our theory compel us to claim, that clamping high is an invariant feature of mental imagery. In other words, our model does not dictate that there be a gradient in change along a *set portion* of the visual hierarchy, only that there exists a gradient in resolution of features encoded below the clamped level, wherever that may be.

With that said, it seems at the present moment that natural imagery generally involves clamping relatively high as indicated by numerous studies finding the greatest similarities between vision and imagery in high-level visual cortex (Pearson et al. 2015). All experiments showing significant similarity between imagery and vision have also shown a high degree of asymmetry (quantified in many studies as the decoding accuracy of a

classifier) in the representations in the early visual cortex (Reddy et al. 2010; Lee et al. 2012). Even in a study that specifically used simple gratings as stimuli, there was decreased performance of a classifier in V1 during imagery relative to vision (Bosch et al. 2014) indicating that even with the simplest stimuli there is still a change in the lowest areas and therefore clamping is likely to generally occur above this level.

Generative vs. Adversarial Imagery

We have interpreted the observed differences between vision and imagery as evidence of feedback from a high-level visual area clamped to an activity pattern that is identical to the mean activity pattern evoked during vision. Let's call this interpretation "generative". Yet we might consider an alternate interpretation in which a subject clamps low but imagines poorly. We will call this the "adversarial" interpretation. Under this interpretation the clamped activity pattern in V1 during imagery of s is not identical to the mean activity pattern evoked while seeing s. Rather, the activity pattern is identical to a blurred or displaced or otherwise corrupted version of s that we'll call s′. This "adversarial image", s′, might get the low-level details of s very wrong while still preserving enough of s that high-level areas can read off the same features they would if s were seen. It is possible that such an arrangement could lead to some, but we believe not all, of the effects we observed in our study. We find the generative interpretation more plausible than the adversarial interpretation for two reasons.

First, the generative interpretation is most parsimonious. As is discussed above, the generative interpretation of "imagining poorly" simply means clamping high (e.g. imagine the correct object "zebra" but fail to imagine the correct frequency of stripes).

The generative interpretation thus replaces the vague notion "imagines poorly" with a single discrete hyperparameter, i.e., the level of clamping. As I described above, clamping high limits spatial resolution and under-specifies features, even if the clamped activity pattern is a perfect reinstatement of a visual activity pattern. The generative interpretation thus attributes limits on the specificity of mental images to the limit of specificity in high-level visual areas that is built-in to the visual system. In contrast, the adversarial interpretation would require additional concepts to model and predict specific imperfections in the reinstatement of cortical activity patterns.

Second and most importantly, the sign of the gradients in signal (i.e., activation amplitude) observed during mental imagery are most compatible with the generative interpretation. Under the generative interpretation signal should attenuate with distance below the clamped level. Signal attenuation with distance from source is a stable configuration for a system–like the visual system–with extensive feedback connections. For the adversarial interpretation to be compatible with our results signal would have to amplify with distance from the clamped level for all levels above in such a way that it can reach activation parity relative to vision near the top of the hierarchy. Such amplification would be unlikely to yield stable dynamics.

Mental imagery and Attention

Previous fMRI studies have shown that changes in signal amplitude, receptive field attributes and feature tuning can be induced by changes in a subject's state of attention (Womelsdorf et al. 2006; Çukur et al. 2013; Klein et al. 2014; Kay et al. 2015; Vo et al. 2017; Klein et al. 2018).In our study subjects were free to marshal attention as needed to

form mental images. It is therefore important to consider how the effects that we have attributed to clamping during mental imagery relate to previously observed attentional effects. Interestingly, from a purely descriptive perspective with regards to the brain areas most affected by imagery or attention, the effects we observed in our imagery study are the opposite of previously observed attention effects. In our study the largest differences between imagery and vision were seen in V1 and V2. These differences weakened with ascent of the visual hierarchy. In contrast several attention studies that reported robust changes in signal amplitude (gain), receptive field size and eccentricity (Klein et al. 2014; Kay et al. 2015; Vo et al. 2017) and/or feature tuning (Çukur et al. 2013) across different states of attention found these changes in high-level visual areas and reported no or smallest changes in V1-V3. Thus it is clear that previously observed attentional effects were not replicated in our study.

The changes in encoding properties induced by changes in attention have often been interpreted as evidence of an attention-induced optimization of the allocation of neural resources. In contrast, we interpret the changes in encoding properties induced by mental imagery as evidence of an inference process that constrains representations of imagined visual features in low-level visual areas to resemble representations of seen features in high-level areas. Whether a similar inference process might be leveraged to explain the disparate effects on representation observed in attention experiments is an interesting topic for future research.

Chapter 6 : Conclusions and Insights Into the Opening Questions

Summary of Dissertation and Results

We tested the hypothesis that mental imagery is a form of inference conditioned on a clamped visual activity pattern at a single processing level in a hierarchical generative model. In Chapter 1, I described the current findings and open questions in the field of mental imagery that we used as a guide in building this hypothesis. In Chapter 2 I described how we developed a formalized treatment of this theory and showed analytically that under this hypothesis activity patterns during imagery are related to activity patterns during vision via an echo transformation, and that the effects of the echo transformation could in principle be revealed by estimating imagery encoding models. Following from these relationships, we predicted that the encoding models would uncover feature tuning in lower visual areas during imagery that closely resembled tuning to seen features in higher areas, as well as a gradient in this distortional effect that increased with hierarchical distance below the clamped area.

In Chapter 3 I described in detail the novel fMRI experiment that we designed and carried out in order to test these predictions, whereby subjects were cued to imagined natural object pictures in different positions of the visual field. I also described the fwRF modeling approach that we used to estimate voxel-wise receptive field and tuning models from data collect during vision runs, and, for the first time, independently from data collected from imagery runs. In Chapter 4 I validate the imagery encoding models by

showing that they can accurately predict activity patterns in response to new stimuli and can be used to decode both the position and content of imagined objects.

By independently estimating separate visual and imagery encoding models for each voxel we were able to directly compare changes in encoded features from vision to imagery across the visual cortex of our subjects. In Chapter 5 I describe the results of this comparison, which corroborated all of the predicted effects, demonstrating relatively larger, more foveal receptive fields and lower spatial frequency tuning during imagery in lower-level visual areas. Additionally, I demonstrated how signal and noise changes from vision to imagery further favored the hierarchical, generative model. In the following sections I discuss the implications of our findings to our understanding of mental imagery by returning to the opening questions.

The Nature of Mental Images and Their Utility

Many aspects of the proposed theory align well with the iconophile stance, and while we may want to be wary of over-projecting our modern understanding of cognitive processes onto the writings of ancient thinkers, the re-instantiation of visual information even appears strikingly similar to Aristotle's "echoes" of visual percepts. On the other hand, the theory we propose here also does not disagree with some aspects of the iconophobe perspective, given that the areas where mental images and percepts seem to be the most interchangeable (at and above the clamped stage) is where the neural code is considered to be more language-like than depictive. By providing evidence for a model of imagery that draws a relationship to vision and emphasizes its representation across a hierarchy, we have presented a portrayal of imagery that is both depictive and descriptive

in nature. Certainly, this will not settle the debate once and for all, but such findings do suggest that we should move away from thinking of imagery within this dichotomy and focus more on its hierarchical representations.

Concerning utility, our theory is consistent with the intuition that mental imagery supports reasoning about things and scenes that are not currently present. Our theory formalizes this intuition by equating the special "reasoning" supported by mental imagery with inference in a hierarchical generative model. The theory thus implicitly asserts that we have mental images because inferring the visual consequences of a predicted or remembered cause can be useful. In visual areas below the clamped area, imagery facilitates inference about the lower-level details associated with, and left unspecified by, the clamped representation. The theory presented here thus views imagery as similar to the phenomenon of "amodal completion" discussed in Revina et al. (2018). It is also consistent with the model of Kosslyn and colleagues (Kosslyn et al. 2006) that treats mental imagery as the faculty that allows one to answer questions such as "Do giraffes have horns?" when no giraffe is handy to inspect. Note that mental imagery is not consistent with any model of vision that treats low and intermediate processing levels (or visual areas) as way-stations in the transformation of images into categorical or propositional representations (e.g., a deep neural network trained to classify objects). Our theory provides a coherent story for why visual information is encoded during imagery in lower areas. In other words, in direct contrast to Pylyshyn's null hypothesis, our theory posits that structured activations in early visual cortex during imagery are not mere

epiphenomenal side effects of ordinary reasoning, but rather reflect the fact that low-level and intermediate representations are in fact worth reasoning about.

If evidence for the theory presented here continues to accumulate it will be interesting to consider a question that the theory currently does not answer: does the visual system support inference because mental images help us to reason, or are we able to generate mental images because inference helps us to see?

How do mental images differ from the ones we see?

In contrast to Hume's depiction of mental images as differing from percepts only in intensity, our results have revealed a significant change in the visual features encoded during mental imagery in the lower parts of the visual hierarchy. Therefore, it appears that visual images are not simply fainter versions of percepts but instead have a more complex relationship to vision: mental images closely resemble percepts in higher areas, but are distorted relative to percepts in terms of the representations they occupy in lower-level areas. This would suggest that the details of imagined objects are difficult to make out not because they are faint, but because the resolution is simply not there.

Interestingly, recent studies have linked the subjective vividness of mental imagery to the similarity of representations in visual cortex during vision and imagery (Dijkstra et al. 2017, 2019). In the theory presented here the similarity between imagery and vision is determined, and crucially limited, by the hierarchical level at which an activity pattern is clamped. Our results suggest that in our experiments clamping occurred at least as high as V4. At and above V4 the distortion of spatial tuning and receptive field attributes

during mental imagery relative to vision were much less than in areas lower in the hierarchy. In summary, within our model, imagery differs from vision (both subjectively and empirically) in that the specific inference it imposes on the visual system results in a distortion in the features encoded below the clamped area.

Vision Synthesis in the Absence of Retinal Input

By tying mental imagery to inference we have provided a potential explanation for how mental imagery could utilize visual representations but encode them in different activity patterns. We have also given empirical support to the intuition that we imagine to "see" the visual consequences of predictions and memories. Our work also extends the power and relevance of the generative perspective on vision. Previous results relating vision to inference have supplied evidence that representations in biological visual systems are adapted to the structure of the visual environment (Olshausen & Field 1996; Karklin & Lewicki 2009; Berkes et al. 2011). Other studies relating vision to the related concept of predictive coding have supplied evidence that knowledge of the visual environment can be combined with contextual information to represent the visual structure of occluded scenes (Muckli et al. 2015) and of illusory contours (De Haas & Schwarzkopf 2018). The current results provide additional compelling evidence that highly structured representations can emerge independently of retinal input (Berkes et al. 2011; Vetter et al. 2014) allowing the visual system to reason coherently about the visual environment even when there is nothing to see.

References

Albers, A.M. et al., 2013. Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, 23(15), pp.1427–1431. Available at: http://dx.doi.org/10.1016/j.cub.2013.05.065.

Albright, T.D., 2012. On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery, and Perception. *Neuron*, 74(2), pp.227–245. Available at: http://dx.doi.org/10.1016/j.neuron.2012.04.001.

Alink, A. et al., 2010. Stimulus Predictability Reduces Responses in Primary Visual Cortex. *Journal of Neuroscience*, 30(8), pp.2960–2966.

American Psychiatric Association, 2013. *Diagnostic and statistical manual of mental disorders* 5th ed., Arlington, VA: American Psychiatric Publishing.

Aristotle, 1984. *The Complete Works of Aristotle, Vol. I and II* J. Barnes, ed., Princeton, NJ: Princeton University Press.

Bar, M., 2009. The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), pp.1235–1243.

Berkes, P. et al., 2011. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013), pp.83–87. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21212356 [Accessed October 2, 2018].

Bosch, S.E. et al., 2014. Reinstatement of Associative Memories in Early Visual Cortex Is Signaled by the Hippocampus. *Journal of Neuroscience*, 34(22), pp.7493–7500. Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0805-14.2014.

Braithwaite, S.R. et al., 2010. The interpersonal theory of suicide. *Psychological Review*, 117(2), pp.575–600.

Brewin, C., 2011. The Nature and Significance of Memory Disturbance in Posttraumatic Stress Disorder. *Ssrn*.

Brewin, C.R. et al., 2010. Intrusive Images in Psychological Disorders: Characteristics, Neural Mechanisms, and Treatment Implications. *Psychological Review*, 117(1), pp.210–232.

Byrne, P., Becker, S. & Burgess, N., 2007. Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, 114(2), pp.340–375.

Carandini, M., 2005. Do We Know What the Early Visual System Does? *Journal of Neuroscience*, 25(46), pp.10577–10597.

Christopher M. Bishop, 2006. *Pattern Recognition and Machine Learning*, Springer.

Cichy, R.M., Heinzle, J. & Haynes, J.D., 2012. Imagery and perception share cortical representations of content and location. *Cerebral Cortex*, 22(2), pp.372–380.

Close, H. et al., 2014. Mental imagery in bipolar affective disorder versus unipolar depression: Investigating cognitions at times of 'positive' mood. *Journal of Affective Disorders*, 166, pp.234–242. Available at: http://dx.doi.org/10.1016/j.jad.2014.05.007.

Coen-Cagli, R., Kohn, A. & Schwartz, O., 2015. Flexible gating of contextual influences in natural vision. *Nature neuroscience*, 18(11), pp.1648–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26436902.

Crane, C. et al., 2012. Suicidal imagery in a previously depressed community sample. *Clinical Psychology and Psychotherapy*, 19(1), pp.57–69.

Çukur, T. et al., 2013. Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6), pp.763–770. Available at: http://dx.doi.org/10.1038/nn.3381.

David, S. V. & Gallant, J.L., 2005. Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, 16(2–3), pp.239–260.

Day, S.J., Holmes, E.A. & Hackmann, A., 2004. Occurrence of imagery and its link with early memories in agoraphobia. *Memory*, 12(4), pp.416–427.

Dennett, D.C., 1969. *Content And Consciousness*, London: Routledge & Kegan Paul.

Dijkstra, N., Bosch, S.E. & van Gerven, M.A.J., 2019. Shared Neural Mechanisms of Visual Perception and Imagery. *Trends in Cognitive Sciences*, xx, pp.1–12. Available at: https://doi.org/10.1016/j.tics.2019.02.004.

Dijkstra, N., Bosch, S.E. & van Gerven, M.A.J., 2017. Vividness of Visual Imagery Depends on the Neural Overlap with Perception in Visual Areas. *The Journal of Neuroscience*, 37(5), pp.1367–1373. Available at: http://www.sebosch.com/wp-content/uploads/2017/03/dijkstra_2017_jon.pdf [Accessed September 25, 2018].

Duke, L.A. et al., 2008. The sensitivity and specificity of flashbacks and nightmares to trauma. *Journal of Anxiety Disorders*, 22(2), pp.319–327.

Dumoulin, S. & Wandell, B., 2008. Population receptive field estimates in human visual

cortex. *NeuroImage*, 39(2), pp.647–660.

Einhäuser, W. & König, P., 2010. Getting real-sensory processing of natural stimuli. *Current opinion in neurobiology*, 20(3), pp.389–95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20434327.

Engel, S.A., Glover, G.H. & Wandell, B.A., 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, 7(2), pp.181–192.

Filgueiras, A., Quintas Conde, E.F. & Hall, C.R., 2018. The neural basis of kinesthetic and visual imagery in sports: an ALE meta − analysis. *Brain Imaging and Behavior*, 12(5), pp.1513–1523.

Freedman, B.J., 1974. The subjective experience of perceptual and cognitive disturbances in schizophrenia. A review of autobiographical accounts. *Archives of general psychiatry*, 30(3), pp.333–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4149762.

Friston, K., 2005. A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1456), pp.815–36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15937014%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1569488.

Gao, J.S. et al., 2015. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9(September), pp.1–12. Available at: http://journal.frontiersin.org/Article/10.3389/fninf.2015.00023/abstract.

Gregory, J.D. et al., 2010. Intrusive memories and images in bipolar disorder. *Behaviour Research and Therapy*, 48(7), pp.698–703. Available at: http://dx.doi.org/10.1016/j.brat.2010.04.005.

Grill-Spector, K. et al., 2018. The functional neuroanatomy of face perception: from brain measurements to deep neural networks. *Interface focus*, 8(4), p.20180013. Available at: https://royalsocietypublishing.org/doi/pdf/10.1098/rsfs.2018.0013.

Grill-Spector, K. & Malach, R., 2004. the Human Visual Cortex. *Annual Review of Neuroscience*, 27(1), pp.649–677.

De Haas, B. & Schwarzkopf, D.S., 2018. Spatially selective responses to Kanizsa and occlusion stimuli in human visual cortex. *Scientific Reports*, 8(1), pp.1–11. Available at: http://dx.doi.org/10.1038/s41598-017-19121-z.

Hackmann, A. & Holmes, E.A., 2004. Reflecting on imagery: A clinical perspective and

overview of the special issue of Memory on mental imagery and memory in psychopathology. *Memory*, 12(4), pp.389–402.

Hansen, K.A., Kay, K.N. & Gallant, J.L., 2007. Topographic Organization in and near Human Visual Area V4. *Journal of Neuroscience*, 27(44), pp.11896–11911. Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.2991-07.2007.

Henriksson, L. et al., 2008. Spatial frequency tuning in human retinotopic visual areas. *Journal of vision*, 8(10), p.5.1-13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19146347.

Hirsch, C.R. & Holmes, E.A., 2007. Mental imagery in anxiety disorders. *Psychiatry*, 6(4), pp.161–165. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1476179307000195.

Holmes, E.A., Geddes, J.R., et al., 2008. Mental imagery as an emotional amplifier: Application to bipolar disorder. *Behaviour Research and Therapy*, 46(12), pp.1251–1258. Available at: http://dx.doi.org/10.1016/j.brat.2008.09.005.

Holmes, E.A. et al., 2019. Mental imagery in psychiatry: conceptual & clinical implications. *CNS Spectrums*, pp.1–13.

Holmes, E.A., Lang, T.J., et al., 2008. Prospective and positive mental imagery deficits in dysphoria. *Behaviour Research and Therapy*, 46(8), pp.976–981.

Holmes, E.A. & Mathews, A., 2010. Mental imagery in emotion and emotional disorders. *Clinical Psychology Review*, 30(3), pp.349–362. Available at: http://dx.doi.org/10.1016/j.cpr.2010.01.001.

Horikawa, T. et al., 2013. Neural decoding of visual imagery during sleep. *Science (New York, N.Y.)*, 340(6132), pp.639–42. Available at: http://content.apa.org/journals/xge/53/5/339 [Accessed April 20, 2015].

Horikawa, T. & Kamitani, Y., 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8.

Hume, D., 2003. *A Treatise of Human Nature*, Mineola, NY: Dover Publications.

Ishai, A. & Sagi, D., 1995. Common mechanisms of visual imagery and perception. *Science*, 268(5218), pp.1772–1774.

Jezzard, P. & Balaban, R.S., 1995. Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic resonance in medicine*, 34(1), pp.65–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7674900.

Just, M.A. et al., 2004. Imagery in sentence comprehension: An fMRI study. *NeuroImage*, 21(1), pp.112–124.

Karklin, Y. & Lewicki, M.S., 2009. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225), pp.83–86. Available at: http://www.nature.com/articles/nature07481 [Accessed October 9, 2018].

Kay, K.N., Winawer, J., et al., 2013. Compressive spatial summation in human visual cortex. *Journal of neurophysiology*, 110(2), pp.481–94. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3727075&tool=pmcentrez&rendertype=abstract [Accessed March 10, 2015].

Kay, K.N., Rokem, A., et al., 2013. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Frontiers in neuroscience*, 7(December), p.247. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3865440&tool=pmcentrez&rendertype=abstract [Accessed March 25, 2015].

Kay, K.N. et al., 2008. Identifying natural images from human brain activity. *Nature*, 452(7185), pp.352–355.

Kay, K.N., Weiner, K.S. & Grill-Spector, K., 2015. Attention Reduces Spatial Uncertainty in Human Ventral Temporal Cortex. *Current biology : CB*, 25(5), pp.595–600. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25702580 [Accessed February 24, 2015].

Kayser, C., Körding, K.P. & König, P., 2004. Processing of complex stimuli and natural scenes in the visual cortex. *Current opinion in neurobiology*, 14(4), pp.468–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15302353.

Kleim, B., Ehlers, A. & Glucksman, E., 2007. Early predictors of chronic post-traumatic stress disorder in assault survivors. *Psychological Medicine*, 37(10), pp.1457–1467.

Klein, B.P. et al., 2018. Cortical depth dependent population receptive field attraction by spatial attention in human V1. *NeuroImage*, 176(April), pp.301–312. Available at: https://doi.org/10.1016/j.neuroimage.2018.04.055.

Klein, B.P., Harvey, B.M. & Dumoulin, S.O., 2014. Attraction of position preference by spatial attention throughout human visual cortex. *Neuron*, 84(1), pp.227–237.

Kosslyn, S.M., Ball, T.M. & Reiser, B.J., 1978. Visual Images Preserve Metric Spatial Information : Evidence from Studies of Image Scanning. , 4(1), pp.47–60.

Kosslyn, S.M. & Thompson, W.L., 2003. When Is Early Visual Cortex Activated during Visual Mental Imagery? *Psychological Bulletin*, 129(5), pp.723–746.

Kosslyn, S.M., Thompson, W.L. & Ganis, G., 2006. *The Case for Mental Imagery*, Oxford University Press. Available at: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195179088.001.0001/acprof-9780195179088.

Kriegeskorte, N., 2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1), pp.417–446. Available at: http://www.annualreviews.org/doi/10.1146/annurev-vision-082114-035447.

Kriegeskorte, N. & Douglas, P.K., 2018. Cognitive computational neuroscience. *Nature Neuroscience*, 21(9), pp.1148–1160. Available at: http://dx.doi.org/10.1038/s41593-018-0210-5.

Lee, S.H., Kravitz, D.J. & Baker, C.I., 2012. Disentangling visual imagery and perception of real-world objects. *NeuroImage*, 59(4), pp.4064–4073. Available at: http://dx.doi.org/10.1016/j.neuroimage.2011.10.055.

Lee, T.S. & Mumford, D., 2003. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America*, 20(7), pp.1434–48. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12868647.

Lipton, M.G. et al., 2010. Distinguishing features of intrusive images in obsessive-compulsive disorder. *Journal of Anxiety Disorders*, 24(8), pp.816–822. Available at: http://dx.doi.org/10.1016/j.janxdis.2010.06.003.

MacKisack, M. et al., 2016. On picturing a candle: The prehistory of imagery science. *Frontiers in Psychology*, 7(APR), pp.1–16.

Markov, N.T. et al., 2013. Cortical high-density counterstream architectures. *Science (New York, N.Y.)*, 342(6158), p.1238406. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24179228.

MCGHIE, A. & CHAPMAN, J., 1961. Disorders of attention and perception in early schizophrenia. *The British journal of medical psychology*, 34(2), pp.103–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/13773940.

Mechelli, A., 2004. Where Bottom-up Meets Top-down: Neuronal Interactions during Perception and Imagery. *Cerebral Cortex*, 14(11), pp.1256–1265. Available at: https://academic.oup.com/cercor/article-abstract/14/11/1256/331439 [Accessed September 25, 2018].

Michael, T. et al., 2005. Unwanted memories of assault: What intrusion characteristics are associated with PTSD? *Behaviour Research and Therapy*, 43(5), pp.613–628.

Moritz, S. et al., 2018. If it is absurd, then why do you do it? The richer the obsessional experience, the more compelling the compulsion. *Clinical Psychology and Psychotherapy*, 25(2), pp.210–216.

Muckli, L. et al., 2015. Contextual Feedback to Superficial Layers of V1. *Current Biology*, 25(20), pp.2690–2695. Available at: http://dx.doi.org/10.1016/j.cub.2015.08.057.

Muller, A.J. et al., 2014. Imagine that: elevated sensory strength of mental imagery in individuals with Parkinson's disease and visual hallucinations. *Proceedings of the Royal Society B: Biological Sciences*, 282(1798), pp.20142047–20142047.

Murray, S.O. et al., 2002. Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23), pp.15164–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=137561&tool=pmcentrez&rendertype=abstract.

Myers, S. et al., 2007. Intrusive images and memories in major depression. *Behaviour Research and Therapy*, 45(11), pp.2573–2580.

Naselaris, T. et al., 2014. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, 105, pp.215–228. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1053811914008428 [Accessed October 31, 2014].

Naselaris, T. et al., 2009. Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63(6), pp.902–915. Available at: http://dx.doi.org/10.1016/j.neuron.2009.09.006.

O'Connor, D.B. et al., 2018. From ideation to action: Differentiating between those who think about suicide and those who attempt suicide in a national study of young adults. *Journal of Affective Disorders*, 241(July), pp.475–483. Available at: https://doi.org/10.1016/j.jad.2018.07.074.

Oertel, V. et al., 2009. Mental imagery vividness as a trait marker across the schizophrenia spectrum. *Psychiatry Research*, 167(1–2), pp.1–11. Available at: http://dx.doi.org/10.1016/j.psychres.2007.12.008.

Olshausen, B.A. & Field, D.J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583), pp.607–609. Available at: http://www.nature.com/doifinder/10.1038/381607a0 [Accessed October 9, 2018].

Osman, S. et al., 2004. Spontaneously occurring images and early memories in people

with body dysmorphic disorder. *Memory*, 12(4), pp.428–436.

Paivio, A., 1963. Learning of Adjective-Noun Paired Associates As a Function of Adjective-Noun Word Order and Noun Abstractness. *Canadian journal of psychology*, 17(4), pp.370–379.

Palmiero, M. et al., 2015. Domain-specificity of creativity: A study on the relationship between visual creativity and visual mental imagery. *Frontiers in Psychology*, 6(DEC), pp.1–8.

Pearson, J. et al., 2015. Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends in Cognitive Sciences*, 19(10), pp.590–602. Available at: http://dx.doi.org/10.1016/j.tics.2015.08.003.

Pearson, J., Clifford, C.W.G. & Tong, F., 2008. The Functional Impact of Mental Imagery on Conscious Perception. *Current Biology*, 18(13), pp.982–986.

Piras, F. et al., 2017. Organization and hierarchy of the human functional brain network lead to a chain-like core. *Scientific Reports*, 7(1), pp.1–13.

Podgorny, P. & Shepard, R.N., 1978. Functional representations common to visual perception and imagination. *Journal of Experimental Psychology: Human Perception and Performance*, 4(1), pp.21–35.

Pylyshyn, Z.W., 2002. Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, 25(02), pp.157–182.

Pylyshyn, Z.W., 2003. *Seeing and Visualizing: It's Not What You Think*, Cambridge, MA: MIT Press.

Pylyshyn, Z.W., 1973. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*.

Rao, R.P.N. & Ballard, D.H., 1999. Predictive coding in the visual cortex : a functional interpretation of some extra-classical receptive-field effects. , pp.79–87.

Reddy, L., Tsuchiya, N. & Serre, T., 2010. Reading the mind's eye: decoding category information during mental imagery. *NeuroImage*, 50(2), pp.818–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20004247.

Reichert, D.P., Seriès, P. & Storkey, A.J., 2013. Charles Bonnet syndrome: evidence for a generative model in the cortex? *PLoS computational biology*, 9(7), p.e1003134. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3715531&tool=pmcentrez&rendertype=abstract.

Reisberg, D., Pearson, D.G. & Kosslyn, S.M., 2003. Intuitions and introspections about imagery: the role of imagery experience in shaping an investigator's theoretical views. *Applied Cognitive Psychology*, 17(2), pp.147–160. Available at: http://doi.wiley.com/10.1002/acp.858.

Revina, Y., Petro, L.S. & Muckli, L., 2018. Cortical feedback signals generalise across different spatial frequencies of feedforward inputs. *NeuroImage*, 180(Pt A), pp.280–290. Available at: https://doi.org/10.1016/j.neuroimage.2017.09.047.

Sack, A.T. et al., 2005. Enhanced vividness of mental imagery as a trait marker of schizophrenia? *Schizophrenia Bulletin*, 31(1), pp.97–104.

Savaki, H.E. & Raos, V., 2019. Action perception and motor imagery: Mental practice of action. *Progress in Neurobiology*. Available at: https://doi.org/10.1016/j.pneurobio.2019.01.007.

Senden, M. et al., 2019. Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception. *Brain Structure and Function*, 0(0), p.0. Available at: http://dx.doi.org/10.1007/s00429-019-01828-6.

Shepard, R.N. & Metzler, J., 1971. Mental rotation of three-dimensional objects. *Science*, 171(3972), pp.701–703.

Shorter, J.M., 2007. VII.—IMAGINATION. *Mind*.

Slotnick, S.D. & Thompson, W.L., 2018. Visual Mental Imagery Induces Retinotopically Organized Activation of Early Visual Areas. , (March).

Snow, M., Coen-Cagli, R. & Schwartz, O., 2017. Adaptation in the visual cortex: a case for probing neuronal populations with natural stimuli. *F1000Research*, 6(0), p.1246. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29034079.

Speckens, A.E.M. et al., 2007. Imagery special issue: Intrusive images and memories of earlier adverse events in patients with obsessive compulsive disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, 38(4), pp.411–422.

Spratling, M.W., 2016. Predictive coding as a model of cognition. *Cognitive Processing*, 17(3), pp.279–305.

St-Yves, G. & Naselaris, T., 2017. The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, (June 2017), pp.1–15. Available at: https://doi.org/10.1016/j.neuroimage.2017.06.035.

Stansbury, D.E., Naselaris, T. & Gallant, J.L., 2013. Natural scene statistics account for

the representation of scene categories in human visual cortex. *Neuron*, 79(5), pp.1025–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23932491 [Accessed November 9, 2013].

Stokes, M. et al., 2009. Top-down activation of shape-specific population codes in visual cortex during mental imagery. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(5), pp.1565–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19193903.

Szpunar, K.K., Watson, J.M. & McDermott, K.B., 2007. Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences*, 104(2), pp.642–647.

Thirion, B. et al., 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4), pp.1104–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17029988 [Accessed January 26, 2015].

Thomas, N.J.T., 2018. Mental imagery. *The Stanford Encyclopedia of Philosophy*, pp.187–191. Available at: https://plato.stanford.edu/archives/spr2018/entries/mental-imagery/.

Vetter, P., Smith, F.W. & Muckli, L., 2014. Decoding sound and imagery content in early visual cortex. *Current Biology*, 24(11), pp.1256–1262. Available at: http://dx.doi.org/10.1016/j.cub.2014.04.020.

Vo, V.A., Sprague, T.C. & Serences, J.T., 2017. Spatial Tuning Shifts Increase the Discriminability and Fidelity of Population Codes in Visual Cortex. *The Journal of Neuroscience*, 37(12), pp.3386–3401. Available at: http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.3484-16.2017.

Wang, L. et al., 2015. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, 25(10), pp.3911–3931.

Wells, A. & Hackmann, A., 1993. Imagery and Core Beliefs in Health Anxiety: Content and Origins. *Behavioural and Cognitive Psychotherapy*, 21(3), pp.265–273.

Wheeler, M.E., Petersen, S.E. & Buckner, R.L., 2000. Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20), pp.11125–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11005879.

Winlove, C.I.P. et al., 2018. The neural correlates of visual imagery: A co-ordinate-based meta-analysis. *Cortex*. Available at: https://www.sciencedirect.com/science/article/pii/S0010945217304227?via%3Dihub [Accessed March 27, 2018].

Winlove, C.I.P. et al., 2018. The neural correlates of visual imagery: A co-ordinate-based meta-analysis. *Cortex; a journal devoted to the study of the nervous system and behavior*, 105, pp.4–25. Available at: https://doi.org/10.1016/j.cortex.2017.12.014.

Womelsdorf, T. et al., 2006. Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nature neuroscience*, 9(9), pp.1156–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16906153.

Xiao, J. et al., 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Yuille, A. & Kersten, D., 2006. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7), pp.301–308.